T.C.
EGE ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

DERİN ÖĞRENME YÖNTEMLERİ İLE MEDİKAL BİLGİ ÇIKARIMI

Azer ÇELİKTEN

Danışman(lar) : Prof. Dr. Hasan BULUT
Prof. Dr. Aytuğ ONAN

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Doktora Programı

İzmir
2024

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

ETİK KURALLARA UYGUNLUK BEYANI

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

12/07/2024

İmzası

Azer Çelikten

ÖZET

DERİN ÖĞRENME YÖNTEMLERİ İLE MEDİKAL BİLGİ ÇIKARIMI

ÇELİKTEN, Azer

Doktora Tezi, Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Prof. Dr. Hasan BULUT

Tez Danışmanı: Prof. Dr. Aytuğ ONAN

12.07.2024, 94 sayfa

Biyomedikal alandaki artan makale sayısıyla birlikte, hastalıklar ve semptomlar hakkında keşfedilen değerli bilgiler akademik literatürde saklı kalmaktadır. Bu tezde, solunum yolu hastalıkları ve semptomları arasındaki ilişkilerin keşfi, NLP'nin bir alt disiplini olan ilişki çıkarımı yöntemleriyle ele alınmıştır. Bu amaçla, hastalık-semptom bilgileri içeren bilimsel tıp makalelerinin özetlerinden oluşan özgün bir veri seti hazırlanmıştır. Tezde, semantik benzerlik ve graf gömme yöntemlere dayalı yaklaşımlar önerilmiştir. Semantik benzerlik yöntemleri (Kosinüs, Öklid ve Nokta Çarpımı) karşılaştırılarak hastalık-semptom ve hastalık-hastalık ilişkilerini doğru bir şekilde tespit edebilen yöntemler belirlenmiştir. Ayrıca, ilişki çıkarımı, bağlantı tahmini problemi olarak ele alınarak, graf embedding tabanlı yöntemlerin (TransE, DistMult, ComplEx, HolE) büyük dil modelleri ile çıkarımı ve doğrulanmasını içeren yenilikçi bir yaklaşım geliştirilmiştir. Potansiyel bağlantıların doğruluğunu arttırmak amacıyla küme üçgenleri ve varlık sıklığı yöntemlerinin hibrit kullanımı önerilmiştir.

Elde edilen sonuçlar semantik benzerlik tabanlı yöntemlerden nokta çarpımı benzerliğinin, graf gömme yöntemlerinden ise TransE ve GPT-4 dil modelinin daha başarılı olduğunu göstermiştir. Bu tezde geliştirilen yöntemler, solunum yolu hastalıkları ve semptomları arasındaki ilişkilerin keşfedilmesiyle klinik karar verme süreçlerine ve tıbbi araştırmalara önemli katkılar sağlama potansiyeline sahiptir.

Anahtar sözcükler: Medikal Bilgi Çıkarımı, Semantik Benzerlik, Graf Embedding, Büyük Dil Modelleri.

ABSTRACT

MEDICAL INFORMATION EXTRACTION USING DEEP LEARNING METHODS

ÇELİKTEN, Azer

Ph.D. in Computer Engineering

Supervisor: Prof. Dr. Hasan BULUT

Supervisor: Prof. Dr. Aytuğ ONAN

12.07.2024, 94 pages

With the growing number of articles in the biomedical field, valuable information about diseases and symptoms discovered remains hidden within academic literature. In this thesis, the discovery of relations between respiratory diseases and symptoms utilized relationship extraction techniques from natural language processing. A unique dataset of scientific medical article abstracts with disease-symptom data was created. The study involved comparing semantic similarity (Cosine, Euclidean, and Dot Product) and graph embedding methods (TransE, DistMult, ComplEx, HolE) integrated with large language models to innovate in disease-symptom relationship prediction.

The results demonstrated that dot product similarity outperformed other semantic similarity-based methods. Additionally, TransE and GPT-4 language models surpassed graph embedding-based methods in effectiveness. The methodologies developed in this thesis hold substantial potential to advance clinical decision-making processes and medical research by elucidating the relationships between respiratory diseases and their associated symptoms.

Keywords: Medical Information Extraction, Semantic Similarity, Graph Embedding, Large Language Models.

ÖNSÖZ

Bu tezde, danışmanlarım ile birlikte yürüttüğümüz çalışmada, medikal bilgi çıkarımı alanındaki önemli bir konu olan hastalık ve semptomlar arasındaki ilişkilerin tespit edilmesine yönelik oldukça sınırlı çalışma gerçekleştirildiğini gördük. Araştırmamızın başlangıç noktası, COVID-19 pandemisi sırasında belirlendi. COVID-19'un zaman içinde değişen yapısı, semptomları ve diğer hastalıklar ile ilişkisinin tespitine katkı sağlamak amacıyla çıkış noktamız COVID-19 hastalığına yönelik medikal bilgi çıkarım yöntemleri geliştirmek oldu. Sonrasında, bronşit, astım ve pulmoner emboli gibi diğer solunum yolu hastalıklarını da çalışmaya dahil ederek tez çalışmasının kapsamını genişlettik.

Bu süreçte, medikal metinler üzerinde derin öğrenme ve NLP tekniklerini uygulayarak çeşitli çözüm yaklaşımları geliştirdik. Sunulan bulguların, klinik süreçlerin iyileştirilmesine önemli katkılar sağlayacağına inanıyoruz. Her bir veri noktasının gerçek dünyadaki bir hastanın yaşamıyla bağlantılı olması, bu çalışmanın önemini ve etkisini daha da vurgulamaktadır.

İZMİR

12/07/2024

Azer Çelikten

İÇİNDEKİLER

Sayfa

ŞEKİLLER DİZİNİ

Şekil **Sayfa**

TABLOLAR DİZİNİ

Çizelge **Sayfa**

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar	Açıklama
NLP	Natural Language Processing (Doğal Dil İşleme)
NER	Named Entity Recognition (Varlık İsmi Tanıma)
PCA	Principal Component Analysis (Temel Bileşen Analizi)
MESH	Medical Subject Headings (Tıbbi Konu Başlıkları)
UMLS	Unified Medical Language System (Birleşik Tıp Dil Sistemi)
LSTM	Long Short Term Memory (Uzun Kısa Süreli Bellek)
Bi-LSTM	Bidirectional Long Short Term Memory (Çift Yönlü Uzun Kısa Süre
CNN	Convolutional Neural Network (Konvolüsyonel Sinir Ağları)
CRF	Conditional Random Fields (Koşullu Rastgele Alanlar)
GNN	Graph Neural Networks (Graf Sinir Ağları)
OBO	Open Biological and Biomedical Ontology (Açık Biyolojik ve Biyom
PALM2	Pathways Language ModeL 2
MR	Mean Rank (Ortalama Sıra)
MRR	Mean Reciprocal Rank (Ortalama Karşılıklı Sıra)
CUI	Context Unique Identifier (Bağlam Benzersiz Tanımlayıcısı)

SİMGELER VE KISALTMALAR DİZİNİ (devam)

Kısaltmalar	Açıklama
CUI	Context Unique Identifier (Bağlam Benzersiz Tanımlayıcısı)
POS	Part of Speech (Sözcük Türü)
LLM	Large Language Model (Büyük Dil Modeli)

1 GİRİŞ

1.1 Problemin Tanımı

Günümüzde tıp alanındaki ilerlemelerle birlikte hastalıklar ve semptomlar hakkında araştırmalar yapılmakta ve yeni bilgiler keşfedilmektedir. Sağlık profesyonelleri ve araştırmacılar tıbbi gelişmeler sonucu elde edilen bu bilgileri bilimsel literatürü barındıran veri tabanları aracılığıyla takip ederler. Bu veritabanları arasında, Ulusal Tıp Kütüphanesi tarafından yönetilen ve biyomedikal literatürün geniş bir koleksiyonunu sunan PubMed öne çıkmaktadır (PUBMED, 2023). PubMed, 36 milyonu aşkın makale alıntısı ve onbinlerce bilimsel yayın içerdiğinden bilimsel araştırmalar için önemli bir kaynaktır (PubMed Overview, 2023). Ancak, PubMed'deki bilgilerin yapılandırılmamış formatı, tıbbi gelişmelerin keşfi, incelenmesi ve analizi için manuel arama yöntemlerinin kullanılmasını gerektirmektedir. 2022 yılında yapılan istatistiklere göre, bilimsel literatürde 2,58 milyar arama gerçekleştirilmiştir (MEDLINE PubMed Production Statistics, 2023). Bu durum, veri analizi ve bilgi yönetimi için otomatikleştirilmiş çözümlerin önemini ortaya koymuştur. Manuel bilgi çıkarımı, büyük veri hacmi nedeniyle zaman alıcı, maliyetli ve verimsiz bir yaklaşımdır. NLP yöntemlerinde ve derin öğrenme mimarilerindeki önemli gelişmeler sonucunda medikal metinlerdeki bilgilerin yapılandırılması, analiz edilmesi ve ilişkilendirilmesi ile ilgili etkili çözümler ortaya çıkmıştır.

1.2 Medikal Bilgi Çıkarımı

Medikal bilgi çıkarımı süreci, biyomedikal metin madenciliği ve NLP tekniklerinin bir araya gelmesiyle gerçekleşir. Temel olarak iki ana adımdan oluşur. İlk adım, hastalıklar, semptomlar, ilaçlar, kimyasallar, proteinler gibi medikal terminolojiyi içeren varlıkların metinlerden otomatik olarak tanımlanmasını içerir. Bu aşama, metin içerisinde belirli terimlerin veya ifadelerin tanınması ve sınıflandırılması işlemini kapsar. Varlık tanıma sürecinde, geleneksel sözlük tabanlı ve kural tabanlı yöntemlerin yanı sıra, son dönemde dil modellerinin de-

rin öğrenme tabanlı yaklaşımları büyük popülerlik kazanmıştır. Derin öğrenme modelleri, metinler içerisinden medikal terimlerin tanınmasında, özellikle de tıbbi ifadelerin karmaşıklığı, eş anlamlı ifadelerin sık kullanılması ve özelleşmiş terminolojisi gibi zorlukların üstesinden gelmede etkili çözümler sunmaktadır.

İkinci adım, ilk adımda tanımlanan varlık isimleri arasındaki spesifik ilişkilerin belirlenmesidir. Bu aşama, belirlenen varlıklar arasındaki bağlantıların, yani ilişkilerin keşfedilmesi ve bu ilişkilerin türlerine göre sınıflandırılmasını içerir. İlişki çıkarımında, denetimli öğrenme yöntemleri, etiketli veri setleri üzerinde modellerin eğitilmesini ve bu modellerin yeni örnekler üzerinde tahmin yapmasını sağlar. Denetimli öğrenme süreçleri uzman bilgisi kullanılarak etiketlenmiş veri setlerini gerektirmektedir. İlişki çıkarımında denetimsiz öğrenme yaklaşımları da kullanılmaktadır. Benzerlik tabanlı öğrenme ve graf tabanlı yöntemler gibi denetimsiz öğrenme yaklaşımları, belirlenen varlıklar arasındaki potansiyel ilişkilerin keşfedilmesi için kullanılır. Özellikle graf tabanlı yöntemler, medikal bilgi grafları aracılığıyla hastalıklar, semptomlar ve ilaçlar gibi medikal varlıklar arasındaki karmaşık ilişkilerin görselleştirilmesi, analiz edilmesi ve potansiyel bağlantıların keşfedilmesi konusunda güçlü araçlar sunar.

1.3 Motivasyon

Bilimsel ve tıbbi gelişmelerin sağlık alanındaki önemi göz önüne alındığında, solunum yolu ve akciğer hastalıklarının yönetimi açısından da kritik öneme sahiptir. Solunum yolu hastalıklarının erken teşhisi, tedavi süreçlerini optimize edebilir ve hasta sağlığı üzerinde olumlu etkiler yaratabilir. Tıptaki gelişmelerle birlikte solunum yolu hastalıkları, belirtileri, tanıları ve tedavileri hakkında yeni bilgiler ortaya çıkmaktadır. Solunum yolu ve akciğer hastalıkları, Avrupa Birliği'ndeki sağlık sorunları içinde ölüm oranları bakımından öne çıkan ciddi bir hastalık grubunu temsil etmektedir. Bu hastalıklar, geniş bir yelpazeyi kapsayarak bireylerin sağlığını ve yaşam kalitesini doğrudan etkilemektedir. Aynı zamanda bu hastalıkların yaygınlığı ve bunlara bağlı mortalitenin fazla olması, hastane yatışlarını, ilaç maliyetlerini ve sağlık harcamalarını arttırarak sağlık sistemleri üzerinde önemli maliyet ve yükler oluşturmaktadır.

Özellikle kronik solunum yolu hastalıkları, akut enfeksiyonlar ve pulmoner komplikasyonların yanı sıra bu hastalıkların bulaşıcılığının yüksek olması toplum sağlığı için önemli riskler oluşturduğu için erken teşhis ile etkin tedavi yöntemlerinin geliştirilmesi ve uygulanması, bu riskleri azaltmada kritik bir öneme sahiptir. Solunum yolu hastalıkları, ateş, öksürük ve nefes darlığı gibi ortak semptomlar sergilemekle birlikte, farklı patofizyolojik mekanizmalara sahiptirler. Bu nedenle, doğru teşhisin konularak uygun tedavi yöntemlerinin uygulanması hayati öneme sahiptir.

Özellikle 2019 yılında ortaya çıkan ve küresel bir sağlık krizine yol açan COVID-19, solunum yolu hastalıklarının spektrumunu genişleterek pulmoner komplikasyon riskini artırmıştır. Bu nedenle, hastalıkların erken evrelerinde doğru tanı koymak sağlık sonuçlarını iyileştirmek için kritik önem taşımaktadır. COVID-19 gibi salgın hastalıkların zaman içinde mutasyona uğraması ve semptomlarının değişebilmesi nedeniyle bilimsel literatür içerisinden bu bilgileri otomatik yöntemlerle tespit edebilmek salgının kontrol altına alınabilmesi, erken teşhis açısından oldukça önemli olmasının yanı sıra bu konularda oluşacak bilgi kaynaklarına hızlı bir şekilde ulaşmak ve yeni gelişmeler ile gerekli güncellemeleri yapabilmek açısından önemlidir. Özetle, hastalıklar ve semptomlar arasındaki ilişkilerin belirlenmesi, erken teşhis, klinik karar destek sistemlerinin geliştirilmesi, biyomedikal bilgi grafikleri oluşturulması ve ontolojilerin ve hastalık epidemiyolojilerinin zenginleştirilmesi gibi alanlara önemli katkı sağlamaktadır.

Medikal NER ve ilişki çıkarımı alanındaki çalışmaların çoğu, araştırmaya açık kaynak veri kümelerinden faydalanılarak ilaç-protein (Sun et al., 2020), protein-protein (Zhou et al., 2019a),(Zhou et al., 2019b) hastalık-kimyasal (Onye et al., 2018) ve ilaç-ilaç (Deng et al., 2020), (Feng et al., 2020) etkileşimlerini tespit etmek için çalışmalar gerçekleştirilmiştir. Hastalık-semptom ilişkilerini belirlemeye yönelik etiketli veri kümesine ulaşmadaki kısıtlar ve hastalık-semptom adlarının birbirinin yerine kullanılabildiğinden ayırt edilebilmesindeki zorluklar nedeniyle kısıtlı sayıda çalışma mevcuttur. COVID-

19 hastalığı ve semptomları üzerinde çalışmalar bulunsa da solunum yolu hastalıkları ve semptomlarını geniş kapsamda konu alan bir çalışma mevcut değildir. Bu nedenle, biyomedikal alanda doğal dil anlama açısından solunum yolu hastalıklarını ve semptomları birbirinden ayırmak ve aralarındaki ilişkiyi belirlemek önemli bir araştırma konusudur.

1.4 Tezin Katkıları

Bu tezde, COVID-19, astım, bronşit ve pulmoner emboli olmak üzere dört adet solunum yolu hastalığı üzerinde NLP teknikleri ile detayli analizler yapılarak hastalık ve semptom ilişki çıkarımı ve alt problemlerine yönelik çözüm yaklaşımları geliştirilmiştir. Tezde, hastalık ve semptom ilişkilerini tespit etmek için iki adet yöntem önerisi sunulmuştur. Bunlar semantik benzerlik yöntemler ve graf tabanlı yöntemlerdir. Semantik benzerlik yaklaşımlarında vektörler arasındaki mesafeyi esas alan kosinüs benzerliğinin yanı sıra nokta çarpımı ve öklid benzerliğinin de hastalık ve semptom ilişkilerinin tespit edilebilmesi üzerindeki etkileri araştırılmıştır. Bu yaklaşımlar ile hastalıklar ile ilişkili olabilecek ancak sağlık kaynaklarında yaygın olarak yer verilmeyen birçok nadir semptom literatürden tespit edilebilmiştir. Graf tabanlı yaklaşımlarda ise, büyük dil modellerinden faydalanılarak hastalık-semptom-kimyasal ilişkilerinden oluşan özgün bir bilgi grafı oluşturulmuştur. Bu bilgi grafındaki potansiyel ilişkilerin tespit edilmesi bağlantı tahmini problemi olarak ele alınmıştır. Elde edilen sonuçlar bilimsel literatür ile doğrulandığında semantik benzerlik ve graf tabanlı yöntemler kullanılarak birçok potansiyel hastalık-semptom ve hastalık-hastalık ilişkisi tespit edilebildiği görülmüştür.

Bu tez, solunum yolu hastalıkları ve ilişkili semptomları üzerine odaklanan kapsamlı ve yenilikçi bir araştırma sunmaktadır. Tez çalışmasının katkılar şu şekilde özetlenebilir:

- 16,194 adet Pubmed makale özetinden oluşan ve COVID-19, bronşit, astım ve pulmoner hastalıkları ve semptomlarını içeren özgün bir veri seti oluşturulmuştur.

- Hastalıklar ve semptomların medikal metinlerden otomatik olarak çıkarılması için ontoloji ve ön eğitimli dil modelinden oluşan hibrit bir NER yöntemi önerilmiştir.
- Semantik benzerlik yöntemlerinden kosinüs benzerliğinin yanı sıra nokta çarpımı ve öklid benzerliğinin de hastalık-semptom arasındaki ilişkileri tespit edebilmesi araştırılmıştır. Bu yöntemlerin karşılaştırmalı başarı analizi sunulmuştur.
- Sağlık kaynaklarında belirtilmeyen ancak hastalıklarla ilişkili olma potansiyeli bilimsel literatür ile doğrulanabilen nadir semptomlar tespit edilmiştir. Bu sayede potansiyel olarak gözden kaçırılabilecek semptomların ortaya çıkarılmasına katkı sağlanmıştır.
- Solunum yolu hastalıklarıi semptomları ve kimyasal terimlerden oluşan ve bu ilişkilerin görselleştirilmesi ve analiz edilmesi için bir araç olabilecek literatür tabanlı özgün bir bilgi grafı oluşturulmuştur.
- Bilgi grafını oluşturmak üzere, hastalık-semptom ilişkisini sınıflandıran transformer tabanlı modeller kullanılarak cümlelerin seçimi optimize edilerek bilgi grafının doğruluğu ve güvenilirliği arttırılmıştır.
- Cümlelerden hastalık-semptom ilişkilerini tespit edebilmek amacıyla büyük dil modelleri (GPT 3.5, GPT4 ve Google LaMDA) kullanılarak bilgi grafı için gereken triple'lar elde edilmiştir.
- Oluşturulan bilgi grafı üzerinde çizge gömü (graf embedding) mimarisine dayalı yöntemler (TransE, ComplEx, DistMult, HolE) kullanılarak potansiyel hastalık-semptom, hastalık-hastalık bağlantıları keşfedilmiş ve bu bağlantıların değerlendirilmesi bilimsel literatür kullanılarak gerçekleştirilmiştir.
- Potansiyel bağlantıları daha doğru tespit edebilmek için hibrit bir yöntem önerilerek mevcut yöntemlerden daha başarılı sonuçlar elde edilmiştir.

Tezin ikinci bölümünde medikal NER ve ilişki çıkarma ile ilgili literatürde yer alan çalışmalar açıklanmıştır. Tezin üçüncü bölümünde, NER ve ilişki çıkarımına ilişkin temel yöntemlere, tezin dördüncü bölümünde tez kapsamında

önerilen yöntemlere, tezin beş ve altıncı bölümünde deneysel çalışmalara ve sonuçlarına yer verilmiştir. Son bölümde ise sonuçlar tartışılmıştır.

2 ÖNCEKİ ÇALIŞMALAR

Bu bölümde, tezin kapsamında ele alınan medikal bilgi çıkarımı problemlerine ilişkin literatürde yer alan çalışmalara yer verilmektedir. İlgili çalışmalar biyomedikal varlık ismi tanıma ve biyomedikal ilişki çıkarımı olarak ikiye ayrılarak aşağıda yer alan alt başlıklarda açıklanmıştır.

2.1 Biyomedikal Varlık İsmi Tanıma

NER, bilgi çıkarmanın bir alt alanı olup metin içerisinde geçen isimlerin önceden belirlenmiş bir kategoriye atanması işlemidir. Medikal alanda NER ise hastalık, semptom, tedavi, gen ve protein gibi isimlerin NLP ve makine öğrenmesine dayalı yöntemlerle tespit edilmesi işlemidir. Metinden varlık isimlerinin tespiti, ilişki çıkarımı, soru-cevap sistemleri, metin özetleme gibi görevler için ön adım oluşturmaktadır. Örneğin, ilişki çıkarımı işleminde önce metinden varlık isimlerinin çıkarılması daha sonra bu varlık isimleri arasındaki ilişkinin belirlenmesi gerekmektedir. Literatürdeki NER çalışmaları sözlük tabanlı, kural tabanlı ve makine öğrenmesi tabanlı yöntemler olarak üç kategoride incelenmiştir (Çelikten et al., 2022).

Biyomedikal NER için ilk yöntemler sözlük tabanlı yaklaşımlara dayanmaktadır. Sözlük tabanlı yöntemleri kullanan NER sistemlerinde, her varlık türü için büyük ad koleksiyonlarından oluşan önceden tanımlanmış sözlükler bulunmaktadır. Bu sözlüklerde bulunma durumuna göre metinlerdeki varlık isimleri tespit edilir. National Library of Medicine tarafından geliştirilen MetaMap (Aronson, 2001), biyomedikal varlık isimlerini tanımak için oluşturulan sözlük tabanlı bir sistemdir. Bu sistemde biyomedikal varlık isimleri UMLS konseptleri ile eşleştirilir. DNorm (Leaman et al., 2013a), biyomedikal literatürde hastalık isimlerini tanımak ve sınıflandırmak için tasarlanmış bir araçtır . DNorm, hastalık isimlerini tanımak için sistemli bir sözlük kullanır ve bu terimleri NCBI'nin hastalık ontolojisi ile eşleştirir. PubTutor (Wei et al., 2013b), biyomedikal literatürdeki varlık tanıma ve normalize etme görevleri için bir araçtır. PubTator, çeşitli biyomedikal varlık

tipleri için önceden eğitilmiş modeller sunar ve bunları GenBank, MeSH, UniProt gibi kaynaklarla eşleştirir. cTakes Apache Software Foundation tarafından geliştirilmiş, açık kaynaklı bir doğal dil işleme sistemi olup, biyomedikal ve klinik metinlerden bilgi çıkarmak için tasarlanmıştır (Savova et al., 2010). UMLS sözlüğünü temel alan bir ontolojiye dayanarak, özellikle klinik notlar, sağlık kayıtları gibi metinlerdeki tıbbi bilgileri (örneğin, hastalık teşhisleri, semptomlar, ilaç isimleri ve prosedürler) tanımlamak, sınıflandırmak ve ilişkilendirmek için kullanılır. Sözlük tabanlı yöntemler, özünde daha yüksek hassasiyetle, ancak yeni varlıkların potansiyeli nedeniyle daha düşük hatırlama oranının zayıflığıyla karakterize edilir. Ayrıca, hastalık ve gen adları gibi yaygın biyolojik varlıklar için nispeten iyi yapılandırılmış sözlükler mevcut olsa da, diğer birçok medikal terim için sözlükler kapsamlı veya yeterli değildir.

Kural tabanlı yaklaşımı kullanan NER sistemlerinde, metin kalıplarına göre manuel olarak tanımlanan birkaç kural aracılığıyla adlandırılmış varlıklar tespit edilir. Bir başka deyişle kural tabanlı yöntemler, yapılandırılmış kural kalıpları aracılığıyla potansiyel varlıkları tanımlar.

(Fukuda et al., 1998) tarafından PROPER adlı kural tabanlı bir NER sistemi önerilmiştir . Bu sistemde protein adlarının esas olarak 'p53' gibi temel terimlerden ve 'reseptör' gibi birleştirilmiş terimlerden oluştuğunu gözlemlediler. Bu nedenle, protein adı kalıpları, protein adı terminolojisi, varlık ifade tarzlarındaki çeşitlilik ve farklı ön ekler/son ekler gibi varlık özellikleri kullanılmıştır. Önerilen sistem, 30 ve 50 özetten oluşan veri setleri üzerinde test edilmiş ve %90'ın üzerinde kesinlik ve duyarlılık oranlarında başarı elde edilmiştir.

(Tamames, 2005) tarafından geliştirilen Text Detective isimli bir başka kural tabanlı sistemde gen isimlerinin tespit edilmiştir. Bu sistemde farklı kategorilerde biyolojik varlıkların birleşimleri geniş çaplı etiketlendiğinden, bu kategorilerin birleşimi ile potansiyel varlık isimlerinin etiketlenmesi sağlanmıştır. Gen mention derlemi üzerinde yapılan testlerde %84 kesinlik ve %71 duyarlılık oranında başarı elde edilmiştir.

Kural tabanlı yöntemler söz konusu olduğunda, önceden tanımlanmış

desenler bir varlık sınıfının belirli metinsel özelliklerine de bağlıdır. Başka bir deyişle, varlığa özel sözlükler ve kalıplar, zaman alan süreçler ve uzman bilgisi gerektirir. Sözlük ve kural tabanlı yöntemlere dayalı geleneksel yaklaşımların çoğu, kapsam ve sağlamlık açısından önemli gelişmeler göstermiştir, ancak büyük ölçüde iyi tanımlanmış sözlüklere ve el yapımı kurallardaki bir dizi kelimeye dayanmaktadır.

Makine Öğrenmesi tabanlı yöntemlerde verilere ait öznitelikler ve veri etiketlerinden oluşan veri kümeleri üzerinde modeller eğitilirek, test verileri üzerinde modellerin performansı ölçülür. Makine öğrenmesine dayalı NER sistemleri, kelime sınıflandırma problemi veya dizi etiketleme problemi olarak formüle edilir. Her iki durumda da etiketler, her biri varlık türünden ve varlık isminin sınırını gösteren bir önekten oluşan etiketler kümesidir. Örneğin, bir metinde geçen iki kelimeden oluşan gen ismi 'B-Gene' ve 'I-Gene' etiketleri ile işaretlenir. B-Gene, gen isminin başlangıcını, I-Gene ise gen isminin içerisindeki kelimeleri belirtir. Performansı etkileyen ana faktörler ML modelleri ve özniteliklerdir. ML tabanlı NER yaklaşımları, göreve özel veri kümelerinde iyi performans gösterir. Biyomedikal alandaki NER sistemlerinde, destek vektör makineleri (SVM'ler), gizli Markov modelleri (HMM'ler) ve CRF'ler gibi çeşitli ML modelleri kullanılmıştır (Kazama et al., 2007), (Zhou, 2004), (McDonald & Pereira, 2005).

(Kazama et al., 2007) tarafından, "kelime özelliği", "konuşmanın bir kısmı", "ön ek", "sonek", "alt dizi" ve "önceki sınıf" gibi bir dizi tasarlanmış öznitelik ile bir SVM modeli oluşturulmuştur. GENIA derlemi ile yapılan performans değerlendirmede, protein isimlerini belirlemede %56.5 F1-skorunda, tüm varlık isimilerini belirlemede %51 oranında başarı elde edilmiştir. (Zhou, 2004) tarafından yapılan çalışmada Hidden Markov Model kullanan biyovarlık tanıyıcı ile GENIA derleminde protein tipinde ve tüm tiplerde %75.8 ve %66.6'lık F1-skor başarısına ulaşmışlardır.

CRF, bütün bir cümlenin optimal NER çözümünü bulmak için Viterbi algoritmasını kullanır ve birçok NER görevinde faydalı olduğu bulunmuştur. Bu nedenle, bu yaklaşım yaygın olarak kullanılmaktadır ve son zamanlar-

daki tekrarlayan sinir ağları bile optimal bir diziyi bulmak için CRF'yi de içermektedir. Ayrıca, birçok mimari, CRF tabanlı sistemleri diğer modellerle birleştiren hibrit yaklaşımlar kullanır. (Luo et al., 2018; Wu et al., 2019a; Wei et al., 2019b). (McDonald & Pereira, 2005) tarafından, metinlerdeki gen ve protein isimlerini tanımlamak için, en yararlı öznitelikleri belirlemede bir indüksiyon sistemi kullanarak CRF tabanlı bir model önerilmiştir. 2500 cümlelik bir test seti ile 7500 MEDLINE cümlesi üzerinde gerçekleştirilen eğitimden sonra, önerilen sistem BioCreative I Gen Mention Tanımlama görevinde %82.4'lük bir F1-skoruna ulaşmıştır.

Makine öğrenimi yöntemlerine dayalı yöntemler, özniteliklerin belirlenmesinde insan emeği gerektirdiği ve öznitelik mühendisliği ile etkili özniteliklerin seçimi sistem performanslarına büyük ölçüde etki ettiği için NER sistemleri açısından sınırlıdır. Sözlük tabanlı, kural tabanlı ve makine öğrenmesine dayalı geleneksel yaklaşımlar çoğunlukla büyük ölçekli sözlüklere, hedefe özel kurallara veya iyi yapılandırılmış derlemlere bağlıdır. NER'e yönelik bu yöntemlerin yerini, el yapımı özelliklerden bağımsız olan derin öğrenme tabanlı yaklaşım almıştır.

Derin Öğrenme tabanlı yöntemler, sınıflandırma problemlerini modellemek için farklı ağ mimarilerini kullanır. Evrişimli sinir ağları (CNN'ler), tekrarlayan sinir ağları (RNN'ler) ve uzun kısa süreli bellek modelleri (LSTM'ler) gibi modeller birçok uygulamada yaygın olarak uygulandığı gibi NER görevi için de kullanılmaktadır. DL tabanlı yöntemler, NER'i girdinin cümle olduğu, kelimelerin gömülü olarak sunulduğu bir diziden diziye mimarisi olarak formüle eder. (Batbaatar & Ryu, 2019), tekrarlayan sinir ağı (RNN) derin öğrenme mimarisini kullanarak Twitter mesajlarından sağlıkla ilgili varlıkları tanıma konusunda yüksek hassasiyet ve hatırlama elde etmek için bir yöntem geliştirmiştir. (Wei et al., 2019b), hastalık adlandırılmış varlık tanıma için koşullu rastgele alanlar ve çift yönlü tekrarlayan sinir ağlarını birleştirerek güçlü performans elde etmiştir. (Scepanovic et al., 2020) sosyal medya metinleri üzerinde hastalık isimlerinin tespiti için bir yöntem geliştirmişlerdir. Varlık ismi çıkarımı için Roberta ve Glove kelime vektör yöntemlerini birlikte kullanarak derin öğrenme

algoritmaları ile NER işlemini gerçekleştirmişlerdir. CADEC(Askapatient), Micromed(Twitter), Reddit (MedRed) verisetleri üzerinde sırasıyla 0.82, 0.72 ve 0.73 F1 skor başarı değerleri elde edilmiştir.

Biyomedikal dil modelleri, tıbbi alandaki özel ihtiyaçları karşılamak üzere tasarlanmış, tıbbi metinlerdeki özel terimleri, kavramları ve ilişkileri tanımlamak için kullanılan önceden eğitilmiş modellerdir. Bu modeller, genel dil modellerinden farklı olarak, tıbbi terminoloji, hastalık isimleri, ilaç adları gibi özeliklere odaklanarak eğitilmişlerdir. Bu sayede, tıbbi metinlerin daha doğru bir şekilde işlenmesine, analiz edilmesine ve anlam çıkarılmasına olanak tanırlar. BioWordVec (Zhang et al., 2019) BioBERT (Lee et al., 2019) gibi dil modellerinin yanı sıra SciSpaCy gibi önceden eğitilmiş birçok biyomedikal dil modelini içeren kütüphaneler mevcuttur. SciSpaCy, biyomedikal ve bilimsel metin işleme için tasarlanmış bir Python kütüphanesidir. Bu kütüphane, İsimlendirilmiş Varlık Tanıma işlevi için özel olarak eğitilmiş modeller sunar. SciSpaCy'nin NER modelleri, farklı biyomedikal ve bilimsel veri kümeleri üzerinde eğitilmiş ve bu alana özgü çeşitli varlık türlerini tanıma yeteneğine sahiptir. Bu NER modelleri, hastalıklar, ilaçlar, tedavi yöntemleri, biyolojik işlemler, anatomik terimler, genler ve proteinler gibi medikal terimleri tanıyabilir. Biyomedikal NER yaklaşımlarına ilişkin özet bilgi Tablo 2.1'de verilmiştir.

2.2 Biyomedikal İlişki Çıkarımı

Bu bölümde, biyomedikal ilişki çıkarımına yönelik literatürdeki çalışmalar incelenmiştir. metinlerden medikal varlık isimleri çıkarıldıktan sonra bu varlıklar arasındaki ilişkilerin belirlenmesi sürecine geçilir. İlişkilerin tespitinde genellikle iki ana yaklaşım benimsenir: Denetimli öğrenme ve denetimsiz öğrenme. Denetimli öğrenmede, ilişkilerin çıkarılması bir sınıflandırma problemi olarak ele alınmaktadır. Varlıklar ve ilişkiler alan uzmanları tarafından etiketlendikten sonra, makine öğrenmesi algoritmaları veya yapay sinir ağları kullanılarak ilişkilerin varlığına göre ikili veya çok sınıflı olarak ilişkiler sınıflandırılır. Denetimsiz öğrenme yaklaşımında, etiketlenmiş ilişkilerin bulunmadığı veri set-

Tablo 2.1: Biyomedikal NER Yaklaşımları ve Referansları

Yaklaşım	Yöntemler	Referanslar
Sözlük Tabanlı	MetaMap, cTAKES, DNorm, PubTator	(Aronson, 2001), (Leaman et al., 2013a), (Wei et al., 2013b), (Savova et al., 2010)
Kural Tabanlı	PROPER, TextDetective	(Fukuda et al., 1998), (Tamames, 2005)
Makine Öğrenmesi Tabanlı	CRF, SVM, HMM	(Kazama et al., 2007), (Zhou, 2004), (McDonald & Pereira, 2005)
Derin Öğrenme Tabanlı	RNN, BiLSTM	(Batbaatar & Ryu, 2019), (Wei et al., 2019b), (Scepanovic et al., 2020)
Biyomedikal Dil Modelleri	BioWordVec, BioBERT, Scispacy	(Zhang et al., 2019), (Lee et al., 2019), (Neumann et al., 2020)

lerinde ilişki çıkarımları gerçekleştirilir. Bu yaklaşımlar, istatistiksel yöntemler, benzerlik tabanlı yöntemler ve graf tabanlı yöntemler olarak gruplanabilir. İstatistiksel yöntemlerde, hastalık ve semptomların birlikte ortaya çıkma olasılığı değerlerine dayalı olarak hastalık ve semptomlar arasındaki ilişkiler belirlenir. Benzerliğe dayalı ilişki çıkarmada, kelime vektörleri arasındaki mesafe, vektörler arasındaki mesafeye göre belirlenen kosinüs benzerliği gibi gömme benzerlik yöntemleri kullanılarak hesaplanır. Bu yöntemler ile elde edilen bulgular literatür incelemesi veya uzman doğrulaması ile validasyona tabi tutulur. Bu yaklaşımlar, istatistiksel metodlar, anlamsal benzerlik analizleri ve graf tabanlı teknikler gibi çeşitli yöntemleri içermektedir. Tıbbi metinlerden semptomların belirlenmesine yönelik çalışmalar, belirli hastalık grupları için oluşturulmuş veri setleri üzerinde yapılmıştır. Mental hastalıkların, kalp hastalıklarının ve COVID-19 hastalığının semptom tespiti ile ilgili aşağıda açıklanan çalışmalar bulunmaktadır.

(Wu et al., 2019a) kural tabanlı ve makine öğrenimi modellerini kullanarak semptomları tanımlamak için zihinsel bozukluklarla ilgili bir elektronik sağlık kayıtları veri seti kullanmıştır.

(Uddin et al., 2022) tarafından yapılan çalışmada, Uzun-Kısa Süreli Bellek yöntemi çevrimiçi halka açık bilgi kanalındaki metinlere uygulanarak depresif semptomlar incelenmiştir.

(Eisman et al., 2020) ve (Leiter et al., 2020), derin öğrenmeyi kullanarak klinik notlarda kalp hastalığının semptomlarını tanımlamışlardır.

Koronavirüs pandemisinin başlamasıyla birlikte, COVID-19 hastalığının semptomlarını anlamak için tıbbi metinlerden yararlanmak amacıyla çeşitli çalışmalar gerçekleştirilmiştir.

(Wang et al., 2021), COVID-19 semptomlarını anlamak ve sınıflandırmak için COVID-19 SignSym adını verdikleri bir modül geliştirmiştir. Bu modül, COVID-19 ile ilişkili sözlükler ve örüntü tabanlı kuralların birleştirilmesini içermektedir. Deneysel çalışmalarını, farklı sağlık kaynaklarından elde edilen klinik metinler üzerinde gerçekleştirmişlerdir. COVID-19'a özgü geliştirilen yöntemle 0.972 gibi yüksek bir F skor elde etmişlerdir.

(Lybarger et al., 2021) COVID-19 hastalığına yönelik olarak 1472 klinik nottan oluşan yeni bir semptom veri seti oluşturmuşlardır. Önerilen yapay sinir ağı yöntemi ile, semptom isimlerini tespit etmede 0.83'lük bir F1 skor değerinde başarı elde edilmiştir.

(Zhou et al., 2014) tarafından gerçekleştirilen çalışmada PubMed bibliyografik kayıtlarını hastalık/semptom ile değerlendirerek kosinüs benzerliğini kullanarak hastalık-semptom ilişkileri çıkarılmıştır.

(Hassan et al., 2015), nadir hastalıkların etiketli bir veri setini kullanarak hastalık-semptom ilişkilerini çıkarmak için bir yöntem geliştirmişlerdir. Cümlelerin sözdizimsel örüntüsünü belirlemek ve sırasıyla hastalık ve semptomlar arasındaki ilişkileri bulmak için örüntü öğrenme ve bağımlılık grafiklerini kullanmışlardır.

(Abulaish & Parwez, 2019), iklime duyarlı hastalıklar için hastalık-hastalık, hastalık-semptom ve semptom-semptom ilişkilerini bulmak için bağımlılık ve sözdizimsel kalıpları kullanmışlardır.

(Zlabinger et al., 2020) tarafından sıralama yöntemlerinin performansını değerlendirmek için bir hastalık semptom koleksiyonu oluşturulmuştur. Has-

talıklar ve semptomlar arasındaki ilişki, birlikte görülme istatistikleri kullanılarak belirlenerek, koleksiyon ile değerlendirilmiştir.

(Wada et al., 2018) Q/A modülünden 120.000 cümleden oluşan veri setinde konvolüsyonel sinir ağı mimarisini kullanarak hastalıklar ve semptomlar arasındaki ilişkileri belirlemişlerdir.

Medikal metinlerden ilişkileri çıkarmak için literatürde kullanılan yöntemlerden biri de bilgi graflarıdır. Biyomedikal bilgi grafları, medikal alandaki çeşitli kaynaklardan elde edilen verilerin entegrasyonu ve yapılandırılması yoluyla oluşturularak çeşitli medikal terimler arasındaki karmaşık ilişkileri standartlaştırılmış bir formatta temsil eden yapılardır. Literatürde biyomedikal bilgi graflarının oluşturulması ve eksik bağlantıların tahmin edilebilmesi için farklı yöntemler geliştirilmiştir. (Pechsiri & Piriyakul, 2022) tarafından gerçekleştirilen çalışmada web belgelerden faydalanılarak bir Disease-Symptom Knowledge Graph - Hastalık-Semptom Bilgi Grafı oluşturulmuştur. Bilgi grafının oluşturulmasında PCA yöntemi kullanılmıştır. Bir başka çalışmada AstraZeneca projesi kapsamında özel ve açık kaynaklı veriler kullanılarak makine öğrenmesi yöntemleri ile bilgi keşfini amaçlayan Biological Insight Knowledge Graph geliştirilmiştir (Geleta et al., 2021). Başka bir çalışmada Kawazaki hastalığına dair bir bilgi grafı oluşturulmuştur (Pechsiri & Piriyakul, 2022). Kawazaki hastalığını konu alan ve bu hastalığa dair tedavi yöntemlerinin geliştirilmesini hedefleyen bu çalışmada semantik işleme platformu olan GraphDB'den faydalanılmıştır. Biyomedikal bilgi grafı oluşturan çalışmaların yanı sıra bu graflardaki potansiyel veya eksik kalmış bağlantıları tahmin etmek üzere yöntemlerin geliştirildiği çalışmalar da mavcuttur. (Gao et al., 2022) tarafından bilgi graflarındaki üçlüleri tahmin etmek için, kelime vektörlerine dayalı bir yöntem olan PTMKG-WE önerilmiştir. Makale, zengin tıbbi verilerin ve ilgili ön bilgilerin mevcudiyetine bağlı olarak yeni üçlüleri tahmin etmek için iki iyileştirme stratejisi içeren bir yaklaş sunmaktadır. (Ebeid et al., 2021) tarafından gerçekleştirilen bir başka çalışmada, ilaç ve hedefleri arasındaki bağlantıların tahmini için top-k benzerlik yöntemi kullanılmıştır. Çalışmanın sonuçları, doğru ilaç keşif sonuçları için farklı veri kaynaklarının entegre edilme-

sinin kritik rolünü, bilgi grafiğinin iyileştirilmesinin ve temsil öğrenimi yoluyla tamamlanmasının etkinliğini ve biyomedikal bilgi grafiği bağlantı tahmini görevlerinde ikili sınıflandırıcıların potansiyel fazlalığını vurgulamaktadır. Bu çalışmaların yanı sıra medikal bilgi graflarından olan UMLS ve SNOMED-CT gibi bilinen ontolojilerden faydalanarak da bilgi grafı oluşturma ve tamamlama çalışmaları mevcuttur (Rossanez et al., 2020; Socrates, 2022). Tablo 2.2'de tıbbi metinlerden ilişki çıkarımına ilişkin özet bilgi verilmiştir.

Tablo 2.2: Tıbbi Metinlerden İlişki Çıkarımı Üzerine Çalışmalar

Referans	**Yöntem**	**İncelenen Konu**
(Wu et al., 2020)	Kural Tabanlı, Makine Öğrenimi	Zihinsel bozukluklarla ilgili elektronik sağlık kayıtlarında semptom tanımlama
(Uddin et al., 2022)	LSTM	Çevrimiçi bilgi kanallarında depresif semptomların incelenmesi
(Eisman et al., 2020; Leiter et al., 2020)	Derin Öğrenme	Klinik notlarda kalp hastalığı semptomlarının tanımlanması
(Wang et al., 2021)	Sözlükler ve Örüntü Tabanlı Kurallar	COVID-19 semptomlarının sınıflandırılması ve anlaşılması
(Lybarger et al., 2021)	Yapay Sinir Ağı	COVID-19 semptom isimlerinin tespiti ve yeni semptom veri seti oluşturulması
(Zhou et al., 2014)	Kosinüs Benzerliği	PubMed kayıtlarında hastalık/semptom ilişkilerinin çıkarılması
(Hassan et al., 2015)	Örüntü Öğrenme, Bağımlılık Grafikleri	Nadir hastalıkların semptom ilişkilerinin çıkarılması
(Abulaish & Parwez, 2019)	Bağımlılık ve Sözdizimsel Kalıplar	İklime duyarlı hastalıklar için hastalık-hastalık ve hastalık-semptom ilişkileri
(Zlabinger et al., 2020)	Birlikte Görülme İstatistikleri	Hastalık semptom koleksiyonunda hastalık ve semptomlar arasındaki ilişkiler
(Wada et al., 2018)	Konvolüsyonel Sinir Ağı	Q/A modülü verilerinde hastalıklar ve semptomlar arasındaki ilişkilerin çıkarılması
(Pechsiri & Piriyakul, 2022)	PCA	Hastalık-Semptom Bilgi Grafiği oluşturulması
(Huang et al., 2021)	GraphDB	Kawasaki hastalığı bilgi grafiği
(Gao et al., 2022)	Kelime Vektörleri	Bilgi grafiklerindeki üçlü tahmini
(Rossanez et al., 2020; Socrates, 2022)	Bilinen Ontolojiler (UMLS, SNOMED-CT)	Bilgi grafiği oluşturma ve tamamlama

Literatür çalışmaları incelendiğinde, COVID-19 gibi belirli solunum yolu hastalıklarına odaklanan mevcut araştırmalara rağmen, metin madenciliği tekniklerinin solunum yolu hastalıklarının semptomlarını genel olarak analiz eden kapsamlı bir çalışması mevcut olmadığı görülmektedir. Bu durum, bu alandaki araştırmalara önemli bir katkı sağlama potansiyeli taşımaktadır.

Mevcut çalışmalar, ilişki çıkarımında genellikle kosinüs benzerliği gibi benzerliğe dayalı yöntemlere odaklanmış, ancak diğer semantik benzerlik tekniklerinin kullanımı yeterince araştırılmamıştır. Ayrıca, çoğu yöntem hastalık ve semptomlar arasındaki ilişkileri önceden tanımlanmış bilgilere dayanarak incelerken, biyomedikal literatürün nadir semptomları da içerebileceği ve bu durumun önemli ilişkileri ortaya çıkarabileceği göz ardı edilmektedir. Bu çalışma, denetimsiz öğrenme yöntemleri kullanılarak hastalık-semptom ilişkilerinin benzerlik temelli analizini gerçekleştirerek, bu ilişkilerin kapsamlı bir değerlendirilmesini sunar. Büyük dil modelleri kullanılarak oluşturulan semptom-hastalık biyomedikal bilgi grafiği, hem yaygın hem de nadir semptomların hastalıklarla olan potansiyel ilişkilerini belirlemekte ve bu ilişkiler literatürde doğrulanmaktadır.

3 TEMEL YÖNTEMLER

Bilgi çıkarımı, büyük miktarlardaki yapılandırılmamış veya yarı yapılandırılmış verilerden önemli, değerli ve yararlı bilgilerin elde edilmesi sürecidir. Biyomedikal metinlerdeki bilgi çıkarımı, hastalıklar, semptomlar, ilaçlar, ilaç yan etkileri, proteinler, genler ve diğer biyolojik türler gibi varlık isimlerini elde etmek ve bu varlıklar arasındaki ilişkileri tespit edebilmek için geliştirilen yöntemleri içerir (Çelikten et al., 2022). Bu bölümde biyomedikal bilgi çıkarımının iki alt alanı olan medikal varlık ismi tanıma ve ilişki çıkarımı ile ilgili kullanılan temel yöntemlere yer verilmiştir. NER için, sözlük/ontoloji, kural tabanlı yaklaşımlar ile makine öğrenmesi ve derin öğrenme yaklaşımları incelenmiştir. İlişki çıkarımı problemi ile ilgili kural tabanlı, derin öğrenme ve graf tabanlı yaklaşımlar incelenmiştir.

3.1 Medikal Varlık İsmi Tanıma Yöntemleri

Medikal NER, medikal metinler içerisindeki hastalık, semptom, protein, gen, tedavi, anatomik bölge gibi ifadelerin ve diğer özelleşmiş terimlerin tanımlanıp kategorize edilmesi işlemidir. Örneğin, bir sağlık raporundaki "penisilin" kelimesi bir "ilaç" olarak, "diyabet" kelimesi bir "hastalık" olarak sınıflandırılabilir. Bu teknik, genellikle NLP yöntemleri kullanılarak otomatik olarak gerçekleştirilerek büyük veri kümelerinden özgün bilgilerin hızlı ve etkili bir şekilde çıkarılmasına olanak tanır. Medikal alandaki NER yaklaşımları, sözlük/ontoloji, kural, makine öğrenmesi, derin öğrenme yaklaşımları ile transfer öğrenme ve ön eğitimli dil modellerinin kullanıldığı yaklaşımlar olarak 5 temel alt başlıkta incelenmiştir.

3.1.1 Sözlük/Ontoloji Tabanlı Yöntemler

Bu yaklaşımlar, önceden tanımlanmış ve yapılandırılmış terimler kümesi veya kavramlar hiyerarşisi kullanır. Bu terimler kümesi "sözlük" olarak adlandırılırken, kavramların ve onların arasındaki ilişkilerin detaylı açıklamasını içeren yapıya "ontoloji" denir. Gen fonksiyonlarını tanımlamak için hazırlanan

gen ontolojisi, hastalıkların standartlaştırılmış bir sınıflandırmasını sunarak hastalıklar arasındaki ilişkiler ile ilgili bilgiler içeren hastalık ontolojisi (*Disease Ontology*), PubMed ve diğer biyomedikal literatür veritabanlarında kullanılan, biyomedikal konuları düzenlemek için bir sistemd olan MESH, farklı sağlık bilgisi kaynaklarından gelen bilgileri entegre eden bir meta-sözlük olan UMLS, tıbbi semptomların sistematik bir şekilde sınıflandırılması ve tanımlanması için tasarlanmış bir ontoloji olan semptom ontolojisi medikal sözlük ve ontolojilerden bazılarıdır.

Sözlük/Ontoloji tabanlı yaklaşımlar şu şekilde çalışır:

- Öncelikle, ilgili biyomedikal alanda kullanılan terimlerin kapsamlı bir listesi hazırlanır veya mevcut ontolojilerden elde edilir. Bu liste, hastalıklar, ilaç isimleri, proteinler gibi spesifik kategorilere ait terimleri içerebilir.
- Bu sözlük veya ontoloji, analiz edilecek metinle karşılaştırılır. Metinde geçen her terim, sözlükteki terimlerle eşleştirilir.
- Eşleştirme sonucunda, her bir terimin hangi kategoriye ait olduğu belirlenir. Örneğin, bir terimin "ilaç" kategorisine mi yoksa "hastalık" kategorisine mi ait olduğu tespit edilir.

Bu yaklaşım ile kesin terim eşleştirmeleri sayesinde yüksek doğruluk oranları elde edilir ve metin içindeki terimler doğrudan sözlükteki terimlerle eşleştirildiğinde, süreç hızlı bir şekilde gerçekleşebilir. Ancak, sözlük veya ontolojinin kapsamı dışında kalan terimler tanınması zordur ve ontolojiye eklenen yeni terimlerin güncellenmesi için sürekli bir bakım gereklidir.

3.1.2 Kural Tabanlı Yöntemler

Bu yaklaşımlar önceden tanımlanmış kurallar ve desenler kullanarak metin içerisinde geçen medikal varlık isimlerini otomatik olarak çıkarmayı hedefler. Bu yaklaşımlar, özellikle spesifik terimlerin ve dilbilgisel yapıların varlığını belirlemek için tasarlanmıştır. Oluşturulan kurallar, metin içindeki spesifik kelimelerin, kelime gruplarının veya cümle yapılarının saptanmasına dayanır. Örneğin, bir ilaç adının genellikle büyük harfle başladığı veya

semptomların belli fiillerle ifade edildiği gibi gözlemlere dayanarak kurallar oluşturulabilir.

Kural tabanlı yaklaşımlar şu şekilde çalışır:

- İlk adım, medikal metinlerde sıkça karşılaşılan terimler ve yapılar için desenlerin tanımlanmasıdır. Bu desenler, özgül kelime sınıfları, kelime kökleri veya sözdizimsel yapıların belirlenmesine dair kuralları içermektedir.
- Sonrasında, dilbilgisel işleme tekniklerini (örneğin, lemmatization, POS tagging) kullanarak metin üzerinde ön işleme yapılarak metnin kurallarla daha efektif bir şekilde eşleştirilmesini sağlar.
- Tanımlanan kurallar, metin üzerinde sistem tarafından otomatik olarak uygulanır. Bu süreçte, metin içindeki ilgili ifadeler kurallarla eşleştirilerek tanımlanır ve uygun şekilde kategorize edilir.

Bu yaklaşım ile iyi tanımlanmış kurallar, özellikle dar alanlarda yüksek doğruluk oranları sunarak karar verme süreçlerinde şeffaflık ve anlaşılabilirlik sağlar. Ayrıca, kurallar ihtiyaçlara ve alanlara göre kolayca özelleştirilebilir. Ancak, yeni örnekler ve varyasyonlar için sürekli olarak kuralların güncellenmesi gerekebileceğinde ölçekleme konusunda problemlere neden olabilir. Ek olarak, oluşturulacak kurallar medikal alandaki dilin karmaşıklığı ve medikal dilin doğal çeşitliliğini tam olarak kapsayamayabilir.

3.1.3 Makine Öğrenmesi Yöntemleri

Makine öğrenmesi tabanlı yöntemlerde, modeller öznitelikler ve etiketler içeren eğitim verileri kullanılarak eğitildikten sonra modellerin performansı test verileri kullanılarak ölçülür. Makine öğrenmesi tabanlı sistemlerde, NER görevi bir dizi etiketleme problemi olarak formüle edilir. Bu yöntemde, bir dizideki tokenlar karşılık gelen etiketlerden B, I veya O ile etiketlenir. Bu sürece IOB etiketleme de denir. Bir hastalık adı birden fazla kelime içerdiğinde, örneğin akciğer kanseri, 'B-hastalık' veya 'I-hastalık' olarak etiketlenir. B-hastalık, hastalık adının başlangıcını temsil eder ve I-hastalık, hastalık adındaki kelimeleri

belirtir. İlgili kelime hastalık adında herhangi bir kelime içermiyorsa, örneğin hastalar, ile, O olarak etiketlenir. Makine öğrenmesi yaklaşımlarında, NER'in performansını etkileyen ana faktörler manuel oluşturulan öznitelikler ve kullanılan algoritmalardır. Derin öğrenme yaklaşımlarında ise, benzer şekilde etiketli veri gerekmesinin yanı sıra özniteliklerin çıkarılması model eğitimi sırasında otomatik olarak gerçekleşir. Bu yaklaşımlarda başarılı sonuçlar elde etmek için uzman bilgisi önemlidir. Bi-LSTM ve CNN gibi yöntemler derin öğrenme yaklaşımı olarak medikal varlık isimlerinin otomatik olarak tanınmasında kullanılan yaygın yöntemlerdendir.

Literatürde yaygın olarak kullanılan ve BiLSTM ve CRF yönteminin birlikte kullanılması ile medikal NER alanında yüksek performanslı modeller oluşturulmasına katkı sağlayan bir model önerilmiştir. Bu mimari, hem geçmiş hem de gelecek bağlam bilgisini dikkate alarak metin içindeki varlıkları tanıma ve sınıflandırma yeteneğine sahiptir. Bu model genelde dizi etiketleme yöntemi sonrasında etiketlenen verilerin BiLSTM mimarisine girdi olarak verilmesi ve son katmanda ise CRF yöntemi ile etiket olasılıklarının hesaplanmasına dayanır.

CRF, NER için yaygın olarak kullanılan bir makine öğrenmesi yöntemidir (Lafferty et al., 2001). Bu yöntemde, bir dizilim içerisindeki her birime bir etiket atanır. Olası etiketler üzerinde bir olasılık dağılımı hesaplar ve en olası etiket dizilimini seçer. Buna göre, CRF modeli $p(y^*|x^*)$ olasılığını hesaplamak üzere geliştirilmiş bir olasılık modeli olarak tanımlanmıştır. Burada, $y^* = y_1, ..., y_n$ olası çıktı etiketlerini ve $x^* = x_1, ..., x_n$ giriş verilerini belirtir.

CRF modeli, aşağıdaki denklem ile gösterilebilir:

$$p_\theta(y|x) = \frac{1}{Z_\theta(x)} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t)\right) \tag{3.1}$$

Burada, Z_ϕ tüm olası etiket dizileri için normalleştirme faktörüdür ve şu şekilde tanımlanır:

$$Z_\theta(x) = \sum_{y' \in Y} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \theta_k f_k(y'_{t-1}, y'_t, x_t)\right) \tag{3.2}$$

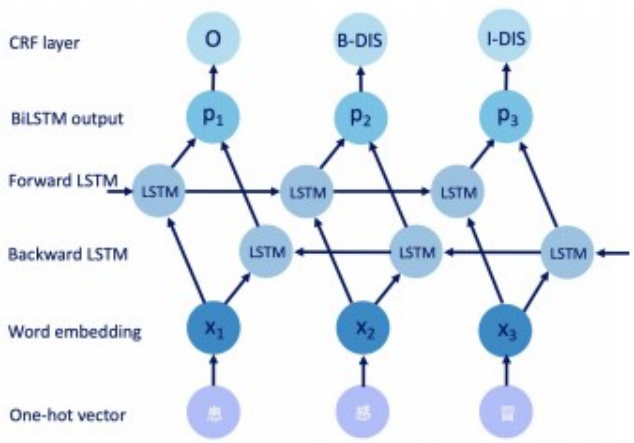
Burada, denklem (3.2)'de de görüleceği üzere, nitelik fonksiyonu paramet-

releri θ_k, t'inci etiket y_t ve $(t-1)$'inci etiket y_{t-1} ve sözcük dizilimi x olan bir fonksiyondur. Nitelik fonksiyonları, makine öğrenmesinde kullanmak istenilen nitelikleri belirleyen fonksiyonlardır. CRF, bütün bir cümlenin optimal NER çözümünü bulmak için Viterbi algoritmasını kullanır. Birçok derin öğrenme mimarisi, CRF tabanlı sistemleri diğer modellerle birleştiren hibrit yaklaşımlar kullanır. Bi-LSTM mimarisi NER görevlerinde CRF ile birlikte kullanılmıştır (Dai et al., 2019).

BiLSTM-CRF modelini kullanarak adlandırılmış varlık tanıma süreci aşağıdaki adımlardan oluşur:

1. Cümledeki her sözcüğün önceden eğitilmiş sözcük yerleştirmeye dayalı sözcük vektörleri modelin ilk katmanına girdi olarak verilir.
2. Modelin ikinci katmanı olan BiLSTM katmanında cümle öznitelikleri otomatik olarak çıkarılır. Bu katman, her bir token için hem geçmiş hem de gelecek bağlamı dikkate alarak bir temsil (*embedding*) oluşturur.
3. Modelin üçüncü katmanı olan CRF katmanına aktarılan kelime temsilleri ile cümlelerdeki etiketlere karar verilir. CRF, en uygun etiket dizisini belirlemek için kelime temsillerini kullanarak tüm dizi boyunca etiketlerin tutarlılığını sağlar.
4. Tüm veriler etiketlenene kadar yukarıdaki 1) ile 3) arasındaki adımları tekrarlanır. Bu yaklaşıma ait mimari Şekil 3.1'de gösterilmiştir. B-DIS ve I-DIS bir hastalık ismini temsil ederken, O simgesi bu hastalık ismi dışında kalan kelimeleri temsil eder.

Medikal NER görevleri için, Makine öğrenimine dayalı yöntemler, özniteliklerin belirlenmesinde insan emeği gerektirdiği ve öznitelik mühendisliği ile etkili özniteliklerin seçimi sistem performanslarına büyük ölçüde etki ettiği için NER sistemleri açısından sınırlıdır. Derin öğrenme yöntemleri geleneksel makine öğrenmesi yöntemlerine kıyasla daha başarılı olmasına rağmen çok miktarda etiketli veri ve donanım kaynağı gerektirdiğinden maliyetlidir.



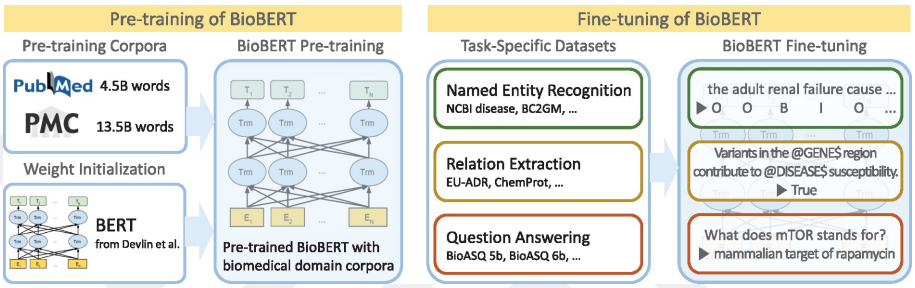
Şekil 3.1: BiLSTM-CRF Mimarisi ile Varlık İsmi Tanıma

3.1.4 Transfer Öğrenme ve Ön Eğitimli Modeller

Transfer öğrenme ve ön eğitimli modeller, biyomedikal NER görevlerinde, özellikle geniş ve karmaşık biyomedikal veri setlerinden özel bilgileri çıkarmak için son yıllarda popülerlik kazanmıştır. Bu yaklaşımlar, genellikle büyük genel amaçlı veri kümeleri üzerinde önceden eğitilmiş ve daha sonra spesifik biyomedikal NER görevleri için ince ayar(*fine tuning*) yapılmış modelleri içerir. Ön eğitimli modeller, özellikle karmaşık dillerin ve terminolojilerin anlaşılmasında yüksek doğruluk oranları sunar. Bu kapsamda BERT tabanlı bir model olan BioBERT ve ön eğitimli NER modelleri bulunduran Scispacy kütüphanesi incelenmiştir.

BERT gibi transformer mimarisini kullanan modeller NER problemleri için kullanılmaktadır. BERT modellerinin medikal ve akademik alan için ön eğitilmiş BioBERT, SciBERT, ClinicalBERT gibi türevleri mevcuttur. Bunlardan en yaygın olan model BioBERT'e ait bilgiler aşağıda verilmiştir. BioBERT, BERT ile neredeyse aynı mimariyi kullanan ve PubMed ve PMC veritabanlarından alınan büyük biyomedikal metinler üzerinde önceden eğitilmiş derin bir bağlamsal modeldir. Şekil 3.2, BioBERT'in ön eğitim sürecini göstermektedir. Biyomedikal alandaki birçok NLP görevi için BERT ve diğer modellerden üstündür. Araştırmaya göre, BioBERT sekiz BNER derleminde ortalama 86,04 F1-ölçüsüne ulaşmıştır: Bunlar, NCBI hastalık, 2010 i2b2/VA,

BC5CDR, BC4CHEMD, BC2GM, JNLPBA, LINNAEUS ve Tür-800 derlemleridir. Ayrıca, BioBERT üç ilişki veri setinde ortalama 81,28 performans gösterir: GAD, EU-ADR, ve CHEMPROT. Genel olarak, BioBERT, bazı veri kümelerinde bazı küçük performans sorunlarına neden olmasına rağmen, önceki en gelişmiş modellerin çoğundan daha iyi performans gösterir. Ancak, BioBERT'in belirli bir NER görevinde etkili bir şekilde performans gösterebilmesi için, görev özgü etiketli veri seti üzerinde ince ayar (*fine-tuning*) yapılması gerekir. İnce ayar sürecinin başarısı, kullanılan etiketli verinin kalitesine ve kapsamına bağlıdır. Yüksek kaliteli ve temsil edici etiketli veri, modelin daha iyi performans göstermesini sağlar.



Şekil 3.2: BioBERT modelinin ön eğitimi

SciSpacy birçok dilde metin işleme görevleri için çeşitli yöntemler sunan Python temelli SpaCy kütüphanesinin biyomedikal metinler üzerinde tekrar eğitilmesi ile geliştirilmiş bir biyomedikal metin işleme aracıdır. NER görevleri için de birçok biyomedikal varlık ismi bu araç sayesinde otomatik tespit edilebilmektedir. Scispacy tarafından sunulan birçok medikal terimi tespit eden modeller bulunmaktadır. Bu modellere ilişkin bilgiler Şekil 3.3'de gösterilmiştir.

Bu modeller arasında yer alan SciSpacy kütüphanesi tarafından sunulan ve hastalık-kimyasal ilişkisi görevi için oluşturulan BC5CDR derlemi ile eğitilen NER aracı `en_ner_bc5cdr_md` ile biyomedikal metinlerdeki hastalık ve kimyasal isimleri tespit edilebilmektedir. Şekil 3.4'te bir PubMed makale özetinde `en_ner_bc5cdr_md` tarafından tespit edilen hastalık isimleri yer almaktadır. Bu modelin birçok hastalık ismini doğru tahmin etmesinin yanı sıra, symptoms, headache, abdominal cramps, diarrhea gibi isimler semptomları ifade ettiği

model	F1	Entity Types
en_ner_craft_md	76.11	GGP, SO, TAXON, CHEBI, GO, CL
en_ner_jnlpba_md	71.62	DNA, CELL_TYPE, CELL_LINE, RNA, PROTEIN
en_ner_bc5cdr_md	84.49	DISEASE, CHEMICAL
en_ner_bionlp13cg_md	77.75	AMINO_ACID, ANATOMICAL_SYSTEM, CANCER, CELL, CELLULAR_COMPONENT, DEVELOPING_ANATOMICAL_STRUCTURE, GENE_OR_GENE_PRODUCT, IMMATERIAL_ANATOMICAL_ENTITY, MULTI-TISSUE_STRUCTURE, ORGAN, ORGANISM, ORGANISM_SUBDIVISION, ORGANISM_SUBSTANCE, PATHOLOGICAL_FORMATION, SIMPLE_CHEMICAL, TISSUE

Şekil 3.3: Scispacy NER Modelleri ve F1-Score Başarıları

halde bu isimlerin de model tarafından hastalık etiketi ile tespit edildiği görülmüştür.

Hastalık İsmi

27086366|t|The Role of TRAF4 and B3GAT1 Gene Expression in the Food Hypersensitivity and Insect Venom Allergy in Mastocytosis

27086366|a|Mastocytosis is an uncommon disease classified as a myeloproliferative neoplasm, however, its symptoms are broad and place patients at crossroads between dermatology, hematology and allergology. Patients with mastocytosis often suffer from symptoms resulting from the activation and release of mediators from the mast cells, such as generalized itching, redness, headache, abdominal cramps, diarrhea, bone pain or arthritis, hypotension, and shock. The possible severe, fatal or near fatal reactions caused by food hypersensitivity are reasons for the research focused on marker identification. The aim of the study was to analyze the gene expression differences in mastocytosis patients with and without food and drug hypersensitivity and insect venom allergy (IVA). A total of 57 Caucasian patients with mastocytosis were studied [median age 41.8; range 18-77 years; 15 (26.3 %) males and 42 (73.7 %) females]. Quantitative RT-PCRs of 11 genes plus ribosomal 18S RNA were run. Symptoms of food hypersensitivity were found in 12 patients (21 %), including 3 patients (13 %) with cutaneous mastocytosis (CM), and 9 (28 %) with indolent systemic mastocytosis (ISM). IVA was confirmed in 13 patients (22.8 %) including 6 patients (10.5 %) with CM, and 7 patients (12.3 %) with ISM. Drug hypersensitivity was diagnosed in 10 patients (17.5 %). Significant differences in the gene expression were found for TRAF4 (p = 0.008) in the comparison of the mastocytosis patients with and without concomitant food hypersensitivity. Furthermore, significant differences were found in gene expression for B3GAT1 (p = 0.003) in patients with IVA compared to patients without insect sting anaphylaxis in the medical history. The expression of studied genes did not differ according to the presence of drug hypersensitivity. The TRAF4 expression was higher in mastocytosis patients with food hypersensitivity in their medical history, the B3GAT1 expression was lower in mastocytosis patients with IVA in history.

Şekil 3.4: Scispacy `en_ner_bc5cdr_md` modeli tarafından tespit edilen hastalık isimleri

3.2 Medikal İlişki Çıkarımı Yöntemleri

Medikal ilişki çıkarımı, tıbbi metinlerden anlamlı bilgilerin otomatik olarak çıkarılması sürecidir. Bu süreç, genellikle NLP tekniklerini kullanarak hastalıklar, semptomlar, ilaçlar ve tedaviler arasındaki ilişkiler gibi önemli tıbbi bilgileri tanımlar ve sınıflandırır. Bu konuda kullanılan yöntemler kural tabanlı, derin öğrenme ve graf tabanlı yaklaşımlar olmak üzere üç alt başlıkta incelenmiştir.

3.2.1 Kural Tabanlı Yöntemler

Medikal ilişki çıkarımı, tıbbi metinlerden anlamlı bilgilerin otomatik olarak çıkarılması sürecidir. Kural tabanlı yöntemler, bu sürecin önemli bir parçasını oluşturur ve metinlerdeki ilişkileri tespit etmek için önceden belirlenmiş kurallar ve desenler kullanır. Kural tabanlı yöntemler, özel bilgileri tanımlamak ve metinlerdeki ilişkileri çıkarmak için çeşitli teknikler kullanır. Bu bölümde, iki yaygın yöntem olan düzenli ifadeler ve sözdizimsel ile semantik kurallar üzerinde durulacaktır.

Düzenli ifadeler, belirli desenlere sahip metin parçalarını tanımlamak ve çıkarmak için kullanılır. Bu yöntem, tıbbi terminoloji, ilaç isimleri ve hastalık adları gibi özel bilgileri tanımlamada oldukça faydalıdır. **Örnek:** "$CRP > 10mg/L$" gibi belirli laboratuvar sonuçlarını tanımlayan düzenli ifadeler, inflamasyon durumunun belirlenmesinde kullanılabilir.

Sözdizimsel ve semantik kurallar, cümle yapısını ve semantik ilişkileri analiz ederek metin içindeki öğeler arasındaki ilişkileri çıkarır. Bu analiz, bağlaçlar, fiiller ve özne/yüklem ilişkilerinin incelenmesini içerir. **Örnek:** "Hasta aspirin alıyor ancak mide ağrısı şikayeti var." cümlesinde, "aspirin" ile "mide ağrısı" arasındaki olası neden-sonuç ilişkisi kurallar aracılığıyla tanımlanabilir.

Kural tabanlı yöntemler, spesifik ve net tanımlanmış durumlar için oldukça etkilidir. Ancak, tıbbi metinlerin karmaşıklığı, bu sistemlerin geliştirilmesi ve

bakımı sırasında zorluklar oluşturabilir. Bu nedenle, kural tabanlı sistemler sıklıkla performanslarını artırmak için diğer yöntemlerle birleştirilir.

3.2.2 Denetimli Öğrenme Yöntemleri

Makine Öğrenimi ve derin öğrenme tabanlı yaklaşımlar, genellikle büyük, etiketlenmiş biyomedikal veri kümelerini (denetimli öğrenme) kullanarak İlişki Çıkarımı işlemleri gerçekleştirmek için kullanılır. Bu veri kümeleri, NLP araçları kullanılarak önceden işlenir ve ardından sınıflandırma modellerini eğitmek için kullanılır. Makine öğrenmesi yöntemlerini iki ana yaklaşıma ayırmak mümkündür: Öznitelik (*feature*) tabanlı ve çekirdek (*kernel*) tabanlı yaklaşımlar. Öznitelik-tabanlı yaklaşımlar, her bir örneği (örneğin, cümle) n-boyutlu bir uzayda bir vektör olarak temsil eder. Destek Vektör Makineleri (SVM) sınıflandırıcıları, genellikle ikili sınıflandırma problemlerini çözmek için kullanılır ve kullanıcının sınıflandırma sürecine müdahalesi olmadığı için "kara kutu" olarak kabul edilir. Bu sınıflandırıcılar, veri özelliklerini temsil etmek için tasarlanmış farklı özellikleri kullanabilir (örneğin, en kısa yol, kelime torbası (BOW) ve POS etiketleme). Çekirdek-tabanlı yaklaşımların temel fikri, bir veri kümesindeki farklı örnekler arasındaki benzerliği, temsillerinin benzerliklerini hesaplayarak belirlemektir. Çekirdek-tabanlı yaklaşımlar, örneklerin yapısal temsilini ekler (örneğin, ayrıştırma ağaçlarını kullanarak). Bu yöntemler, tek bir çekirdek veya çekirdeklerin bir kombinasyonunu kullanabilir.

Derin öğrenme yöntemlerinde ilişki çıkarımı gibi metin madenciliği görevleri için veri temsilleri ile birlikte yapay sinir ağlarını kullanmaktadır. RNN metin madenciliği görevlerinde yaygın olarak kullanılan bir derin öğrenme mimarisidir. RNN, düğümler arasındaki bağlantıların zamansal bir sırayı takip edebildiği bir tür yapay sinir ağı olduğundan her giriş dizisini işlemek için ağ hafızası kullanabilir. Derin öğrenme teknikleri, RNN gibi, kelime gömülmesi, kelime türü etiketleme (*Part of Speech*) ve diğer özelliklere dayalı sınıflandırma modelleri eğitmeyi amaçlar. RNN sınıflandırıcılarının çok katmanlı mimarileri vardır, her katman giriş verilerinin farklı bir temsilini öğrenir. Bu özellik, RNN sınıflandırıcılarını, görev özelikli özellik mühendisliği gerektirmeden, çoklu

metin madenciliği görevlerine uygulanabilir hale getirir.

LSTM ağları, düzenli RNN'e bir alternatiftir. LSTM'ler, uzun bağımlılıkları (örneğin, cümleler) ele alabilen bir RNN türü olduğundan genellikle uzun ve tanımlayıcı olan biyomedikal alan için daha uygundur. İki yönlü LSTM'ler (BiLSTM), her adımda, biri cümleyi sağdan sola, diğeri ise soldan sağa okuyan iki LSTM katmanı kullanır. Her iki katmanın birleşik çıktısı, her adım için nihai bir puan üretir. İki yönlü LSTM'ler, aynı veri kümelerine uygulandığında, geleneksel LSTM'lere göre daha iyi sonuçlar elde etmiştir (Sousa et al., 2020).

3.2.3 Graf Tabanlı Yöntemler

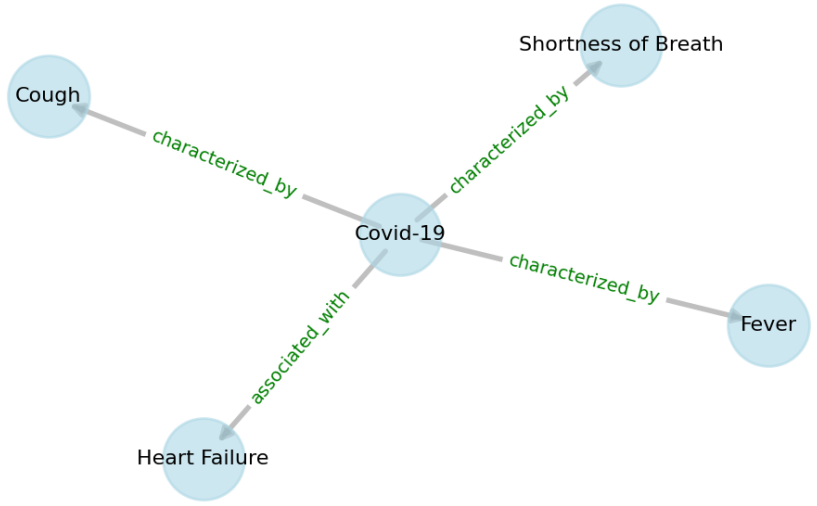
İlişki çıkarımı alanında kullanılan yöntemlerden biri de bilgi grafları (*knowledge graph*)'dır. Bilgi grafları farklı veri kaynaklarını entegre etmenin ve arama gibi uygulamalar için temel ilişkileri modellemenin etkili bir yolu olarak ortaya çıkmıştır. Bu sayede, tıbbi bilgi ve ilişkiler düğümler (*nodes*) ve kenarlar (*edges*) gibi graf yapıları kullanarak modellenir. KG'ler üzerinde kullanılan algoritmalar (en kısa yol, benzerlik, embedding yöntemler) ve gerçekleştirilen analizler ile medikal bilgiler arasındaki ilişkiler çıkarılabilir. Bilgi grafından embedding formatında çıkarılan bilgiler aramayı iyileştirmek, öneri yapmak ve eksik bilgileri tahmin etmek için kullanılır.

Biyomedikal bilgi grafları, her biri belirli ilişki türüyle bağlı olan iki ilgili varlık ve aralarındaki ilişkiyi ifade eden üçlüler (*triples*) kullanılarak yapılandırılır. Örneğin, COVID-19 hastalığı ve ilişkilerini ifade eden üçlüler şu şekilde olabilir:

- COVID-19,characterized_by,fever
- COVID-19,characterized_by,cough
- COVID-19,characterized_by,shortness of breath
- COVID-19,associated_with,heart failure

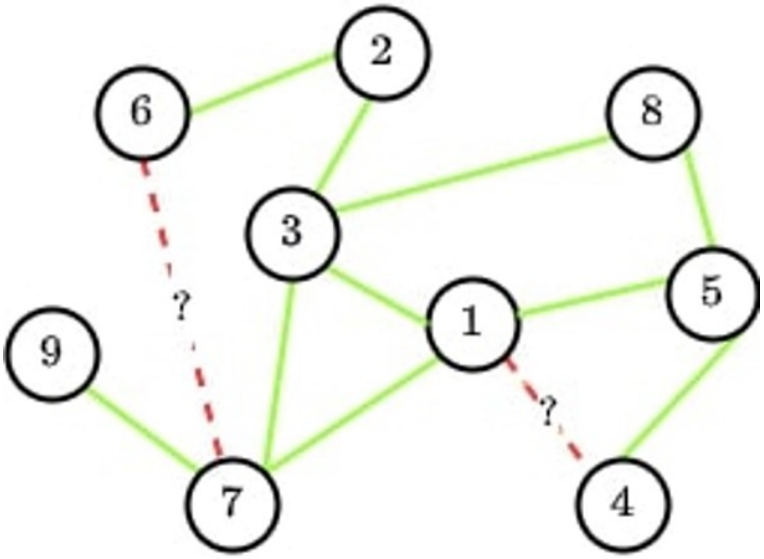
COVID-19'un ateş, öksürük ve nefes darlığı ile karakterize edildiği ve kalp yetmezliği ile ilişkilendirildiği şeklindeki ilişki üçlülerinin bilgi grafiği üzerindeki gösterimi Şekil 3.5'te verilmiştir. Büyük medikal grafların işlenmesi

ve analizi, önemli hesaplama kaynakları gerektirmektedir. Grafların doğruluğu ve işlevselliği kullanılan veri kaynaklarının kalitesine ve yeterli miktarda olmasına bağlıdır.



Şekil 3.5: Örnek Bilgi Grafı Gösterimi

Eksik Bağlantı Tahmini: Bilgi grafları gerçek dünya verilerinin doğası gereği eksik veya hatalı olabilir. Bilgi graflarında eksik bağlantı tahmini (*missing link prediction*), bu eksiklikleri tespit edip doldurmayı amaçlar, böylece daha tam ve doğru bilgi yapısı elde edilir. Eksik bağlantı tahmini, bir bilgi grafiğindeki mevcut verilere dayanarak eksik olan ilişkileri tahmin etme sürecidir. Bu, örneğin, iki ilaç arasındaki potansiyel etkileşimler veya bir hastalığın belirli bir semptomla ilişkilendirilip ilişkilendirilmediği gibi bilgileri içerebilir. Örneğin, bir araştırma alanında iki A ve B kavramı arasındaki bir ilişki bilinirken, bir diğerinde B ve C kavramları arasındaki başka bir ilişki bilinmekte ve A ve C kavramları arasında potansiyel bir ilişki önerilebilir. Eksik bağlantı tahmini problemi Şekil 3.6 gösterilmiştir.



Şekil 3.6: Eksik Bağlantı Tahmini Problemi

Eksik bağlantı tahmini, bilgi graflarını daha güçlü ve yararlı hale getirmenin önemli bir yoludur. Bu yaklaşım, tıptan finansa, birçok farklı alanda uygulama potansiyeline sahiptir. Eksik bağlantı tahmini için graf temsilli öğrenme ve çeşitli bağlantı tahmin algoritmaları kullanılmaktadır. Bu yöntemler aşağıda açıklanmıştır.

Graf Temsili Öğrenme: Grafik yapısındaki verilerin düşük boyutlu, yoğun vektörler olarak temsil edilmesini amaçlayan bir makine öğrenimi yaklaşımıdır. Bu temsiller, genellikle grafikteki düğümler, kenarlar veya daha büyük alt grafikler ile ilişkilendirilir. Amacı, grafikteki karmaşık yapıları ve ilişkileri, algoritmaların daha kolay işleyebileceği sıkıştırılmış bir forma dönüştürmektir. Yakınlık tabanlı yöntemler, rastgele yürüyüş yöntemleri ve GNN gibi yöntemler düğümler ve kenarlar arasındaki çeşitli özellikler dikkate alarak graf temsillerini oluşturmaktadır. Sosyal ağlar, moleküler yapılar, bilgi grafları ve ulaşım ağları gibi alanlarda bu yaklaşım kullanılır.

Bağlantı Tahmin Algoritmaları: Varlık çiftleri arasındaki olası ilişkileri tahmin etmek için çeşitli istatistiksel ve makine öğrenimine dayalı teknikler kullanılır. Bu yöntemler aşağıda açıklanmıştır.

1. İstatistiksel Yöntemler: Bu yöntemler, varlık çiftlerinin birbiriyle bağlantılı

olma olasılığını hesaplamak için düğümler arasındaki yapısal özellikleri kullanır. Örneğin, iki düğüm arasında ortak komşuların sayısı, bu düğümler arasında bir kenarın olasılığı, iki düğümün komşularının kesişim kümesinin, birleşim kümesine oranını ifade eden jaccard katsayısı veya daha az komşusu olan düğümlere daha az ağırlık vermek gibi çeşitli ağırlıklandırma yöntemleri kullanılabilir.

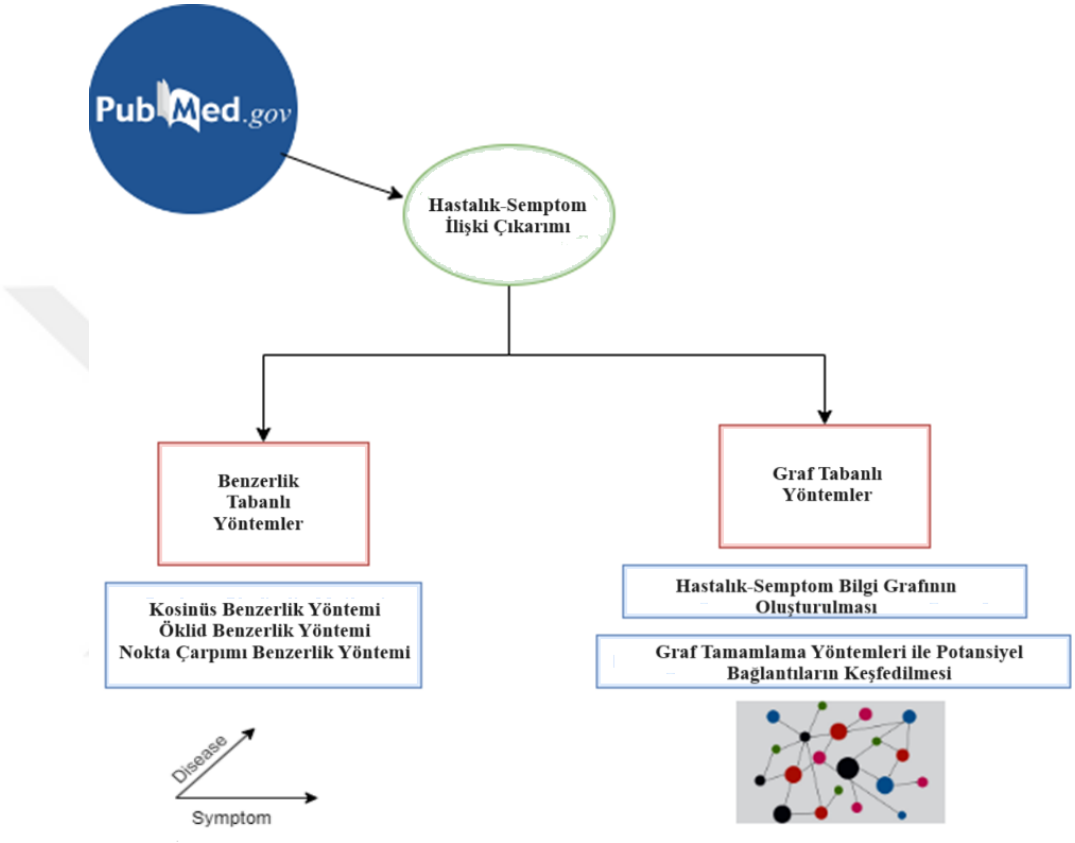
2. Graf Tabanlı Makine Öğrenimi Yöntemleri: Makine öğrenimine dayalı yaklaşımlar, özellik mühendisliği veya otomatik özellik öğrenimi (derin öğrenme yoluyla) kullanarak, daha karmaşık bağlantı tahmin modelleri oluşturur. Önceden etiketlenmiş veriler üzerinde eğitilen makine öğrenmesi modeli yeni varlık çiftleri için bağlantı olup olmadığını tahmini yapar. Bununla birlikte düğümleri vektörler olarak temsil ederek ve bu gömmeleri kullanarak olası bağlantıları tahmin eden yöntemler bulunmaktadır. Node2Vec, DeepWalk, GNN gibi yöntemler bu kategoriye girer. GNN, graf yapılarından otomatik olarak özellikler çıkartabilir ve bu özellikleri bağlantı tahmininde kullanabilir.

TransE, DistMult, ComplEx ve HolE gibi yöntemler de bilgi graflarında ilişki tahmin etme ve varlık gömme (entity embedding) işlemleri için kullanılan modelleme teknikleridir. Bu yöntemler, özellikle ilişkisel verilerin ve bilgi graflarının semantik temsillerini öğrenmekte kullanılan Grafik Gömme (Graph Embedding) veya daha spesifik olarak İlişkisel Gömme (*Relational Embedding*) tekniklerini içermektedir. Bu modeller, grafiklerdeki düğümleri ve kenarları, düğüm ve ilişkilerin özelliklerini kodlayan düşük boyutlu vektörler halinde göstermeyi amaçlar.

4 ÖNERİLEN YÖNTEMLER

Bu bölümde tez kapsamında ele alınan medikal bilgi çıkarımına yönelik önerilen yöntemler açıklanmıştır. Solunum yolu hastalıkları ile ilgili bilgi çıkarımı için bilimsel literatürde yer alan makale özetleri veriseti olarak kullanılmıştır. Bu verilerden anlamlı bilgilerin çıkarılması ve ilişkilendirilmesi

iki temel adımda gerçekleştirilmiştir. İlk olarak metinlerden hastalıklar, semptomlar ve kimyasal varlık isimlerinin otomatik tespit edilmesine yönelik bir yöntem önerilmiştir. Önerilen yöntem alt bölüm 4.1'de açıklanmıştır. Elde edilen varlık isimleri ve aralarındaki ilişkilerin belirlenmesinde benzerlik tabanlı ve graf embedding tabanlı yöntemler kullanılmıştır. Tez çalışmasına ilişkin genel yöntem şeması Şekil 4.1'de verilmiştir.



Şekil 4.1: Genel Yöntem Şeması

Benzerlik tabanlı yöntemler kapsamında ilişki çıkarımı için kullanılan kosinüs, öklid ve nokta çarpımı yöntemleri ile ilgili bilgiler alt bölümlerde verilmiştir.

Graf embedding tabanlı yöntemler kapsamında, bölüm 4.3'te genel bilgi verilmiştir. Bölüm 4.4'te oluşturulan hastalık-semptom bilgi grafı için üçlü üretme adımında kullanılan büyük dil modelleri, Bölüm 4.5'te kullanılan graf embedding yöntemleri ile ilgili bilgi verilmiştir.

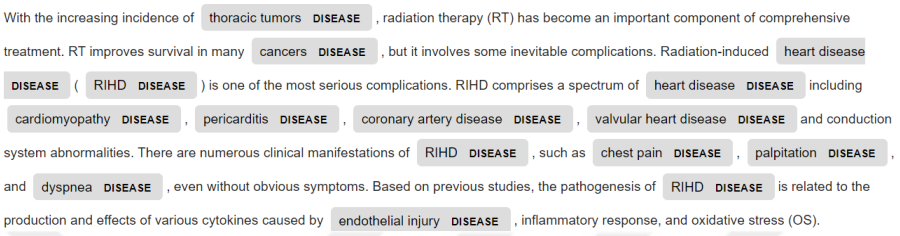
4.1 Medikal Varlık İsmi Tanıma Yöntemi

Hastalıklar ve semptomların tespit edilebilmesi ontolojilere dayalı yöntemler, etiketli veriler üzerinden denetimli öğrenme algoritmaları veya ön eğitimli dil modellerinin kullanımı ile mümkündür. Bu tezde ele alınan solunum yolu ve hastalıklarına ilişkin yeterli miktarda etiketli veri olmadığından denetimli öğrenme algoritmaları yerine ön eğitimli dil modelleri ve ontolojilerden hibrit olarak faydalanılmıştır. Literatürde hastalıklar ile ilgili birçok NER yöntemi olmasına rağmen semptomların belirlenmesine yönelik yöntemler yetersiz kalmaktadır. Bunun en büyük nedeni bir çok bağlamda semptomların da hastalık yerine kullanılabilmesidir. Ancak, bu çalışmada semptom ve hastalıkların birbirleri ile ilişkileri ele alındığından semptom ve hastalık isimlerinin ayrı ayrı tespit edilmesi gerekmektedir. Bu kapsamda önerilen yöntem Scispacy modeli ile varlık isimlerinin tespitinin ardından, bu modeli semptom ontolojisi ile entegre etmek ve elde edilen sonuçlara UMLS normalizasyonu uygulayarak optimize edilmesini sağlamaktadır.

1. Scispacy: SciSpacy, tıbbi ve biyomedikal metinler üzerinde çalışmak üzere tasarlanmış bir Python paketidir. `en_ner_bc5cdr_md` modeli SciSpacy'nin sağladığı medikal NER modellerinden biridir. Bu model, özellikle hastalık ve kimyasal isimlerini tanıma üzerine eğitilmiştir. BC5CDR, BioCreative V Chemical Disease Relation görevi için hazırlanmış bir veri setidir. Bu model, bu spesifik veri seti üzerinde ön eğitilerek bu tür varlıkları tespit etmek üzere optimize edilmiştir. Model, hastalıklar ve kimyasallarla ilgili metinleri analiz etmek için karmaşık öznitelik mühendisliğine gerek duymadan NER görevini gerçekleştirmektedir. Bu modelde DISEASE ve CHEMICAL olmak üzere iki ayrı etiket bulunmaktadır.

   PubMed Merkezi Açık Erişim Alt Kümesi gibi zengin bir biyomedikal metin korpusu üzerinde eğitilmiş word2vec vektörlerine dayanana bu model, 200 boyutlu 50.000 benzersiz kelime vektörü içermektedir. Kimyasal ve hastalık varlıklarını tanıma konusunda 84.28 F1 puanına sahiptir. Bu modelin örnek bir metin üzerinde uygulanması Şekil 4.2'de

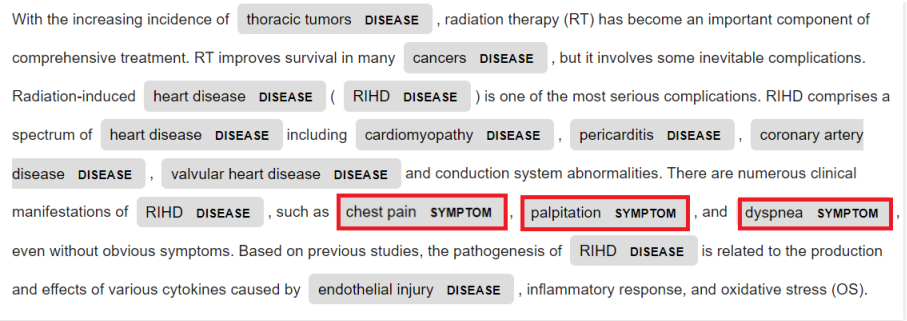
gösterilmiştir. Hastalık isimleri olarak belirlenen "chest pain" (göğüs ağrısı), "palpitation" (çarpıntı) ve "dyspnea" (nefes darlığı) gibi medikal terimlerin aslında belirli bir hastalığın ismini değil, hastalığın belirtilerini ifade ettiği görülmektedir. Hastalıklara ait belirtileri ifade eden bu tür isimlerin semptom olarak tespit edilebilmesi için bu model kullanılırken Semptom Ontolojisi ile entegre edilmiştir. Böyle hastalıklar ve semptomların etiketleri birbirlerinden ayrılabilmiştir.



Şekil 4.2: `en_ner_bc5cdr_md` modeli tarafından tespit edilen hastalık isimleri (Metin: 31592122'den alınmıştır.)

2. Semptom Ontolojisi: OBO Foundry tarafından geliştirilen semptom ontolojisi, klinik gözlemlerin ve hastalık belirtilerinin standartlaştırılması ve sınıflandırılmasını amaçlayan, hem bilimsel doğruluğu hem de mantıksal tutarlılığı ön planda tutan bir ontoloji setidir. Bu ontolojiler ailesi, çeşitli biyomedikal disiplinlerde entegrasyonu ve işbirliğini desteklemek için bir araya getirilmiştir. Semptom ontolojisi ise bu geniş kapsamlı ontoloji setinin önemli bir parçasıdır ve bir hastalık ya da rahatsızlığın varlığını gösteren işlevsel, duyusal ya da görünüşteki değişiklikleri tanımlamak için kullanılır. Semptom ontolojisi, bir hastanın yaşadığı ve hastalığın bir göstergesi olarak rapor ettiği işlevsel, duyusal ya da görünümdeki değişiklikleri gösterir. Bu ontoloji, göğüs ağrısı, kalp çarpıntısı (*palpilation*), nefes darlığı (*dyspnea*) gibi, hasta tarafından hissedilen ve rapor edilen semptomların yanı sıra bu semptomların olası nedenlerini, ilişkili oldukları hastalık durumlarını ve semptomların klinik seyrini de içerir. Toplamda 948 farklı semptom ve bunların ayrıntılı tanımlarını barındıran bu ontoloji, hem araştırmacılar hem de klinisyenler için değerli bir kaynaktır. Hastalık belirtilerinin doğru bir şekilde sınıflandırılması ve

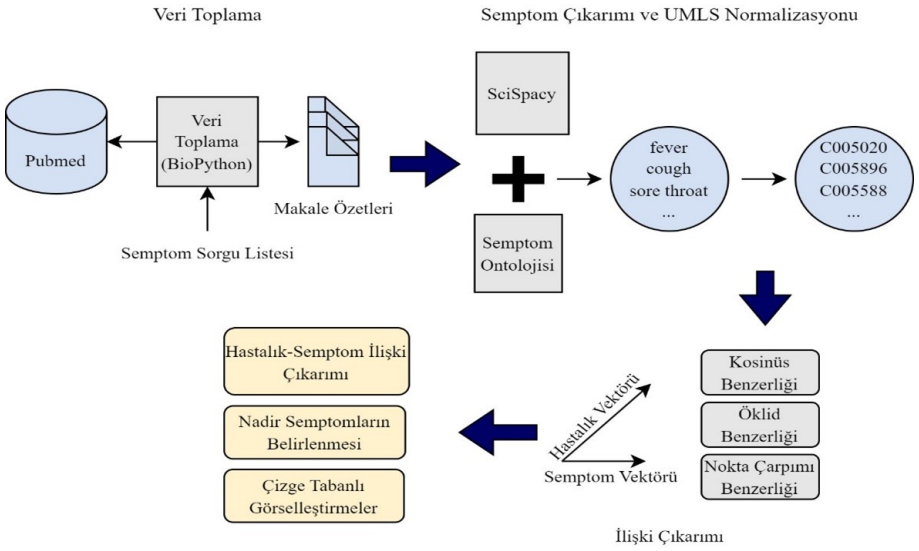
tanımlanması, hastalıkların teşhis ve tedavisinde önemli bir rol oynar. OBO Foundry'nin sağladığı bu ontoloji, sağlık bilgi sistemlerindeki veri değişimini kolaylaştırırken, aynı zamanda farklı çalışma alanlarındaki bilim insanlarının ortak bir dil kullanarak iletişim kurmasına olanak tanır. Scispacy modeli ile elde edilen sonuçlar semptom ontolojisi ile hibrit kullanıldığında hastalık ve semptom isimleri daha doğru bir şekilde ayrıştırılabilmiştir. Semptom ontolojisi entegrasyonu sonrasında NER modelinden elde edilen sonuçlar Şekil 4.3'de gösterilmiştir.



Şekil 4.3: Semptom ontolojisi ve `en_ner_bc5cdr_md` modeli tarafından tespit edilen hastalık/semptom isimleri (Metin PMID: 31592122'den alınmıştır.)

4.2 Semantik Benzerlik Yöntemleri ile İlişki Çıkarımı

Makale özetlerinden elde edilen semptomlar ve hastalıklar arasındaki ilişkilerin tespiti için ScispaCy tarafından sunulan, PubMed ve PMC metinleri üzerinde önceden eğitilmiş ve tıbbi alanın terimlerini etkili bir şekilde temsil eden öğrenilmiş embedding modellerinden faydalanılmıştır. Semptomlar ile hastalıklar arasındaki benzerliği değerlendirmek amacıyla, tüm semptomlar ve hastalıklar vektör uzayında temsil edilerek, aralarındaki mesafe (benzerlik) hesaplanmıştır. Bu vektörler arasındaki benzerliği ölçmek için, gömme benzerlik teknikleri kullanılmıştır. Bu teknikler, vektörler arasındaki mesafeyi hesaplayarak benzerlik ölçüsü sunmaktadır. Daha küçük bir mesafe, vektörlerin (kelimelerin) daha büyük bir benzerlik taşıdığını ifade eder. Semantik benzerlik yönteminin akış şeması Şekil 4.4'de gösterilmiştir.



Şekil 4.4: Benzerlik Tabanlı Hastalık-Semptom İlişki Çıkarımı Yöntemi

Çalışma kapsamında, üç ayrı benzerlik yöntemi kullanılmıştır: Kosinüs, Öklidyen ve nokta çarpım benzerliği. Kosinüs benzerliği, vektörler arasındaki açıyı hesaplayarak iki vektör arasındaki benzerliği ölçer. Öklidyen, vektörler arasındaki doğrudan uzaklığı hesaplar. Nokta çarpım benzerliği ise bir vektörün diğerine olan yansımasını ölçer. Bu yöntemler aracılığıyla, hastalıklar ile semptom vektörleri arasındaki mesafeler hesaplanarak semantik benzerlik temelli ilişkiler açığa çıkarılmaktadır.

Kosinüs Benzerliği (*Cosinus Similarity*): Kosinüs benzerliği, iki vektörün yönleri arasındaki açıyı ölçerek benzerlik derecesini belirler. Bu ölçüm, özellikle kelime vektörlerinin semantik benzerliklerini değerlendirmek için kullanılır. Kosinüs benzerliği, iki vektör arasındaki iç çarpımının, vektörlerin normlarının çarpımına oranı olarak hesaplanır ve 0 ile 1 arasında bir değer alır. Formülü:

$$\text{Kosinüs Benzerliği} = \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{4.1}$$

Öklidyen Benzerliği (*Euclidean Similarity*): İki vektör arasındaki Öklidyen uzaklığa dayalı olarak hesaplanır ve iki vektörün birbirine ne kadar yakın veya uzak olduğunu gösterir. Formülü:

$$\text{Öklidyen Benzerliği} = \frac{1}{1 + \text{Öklidyen Uzaklık}} \tag{4.2}$$

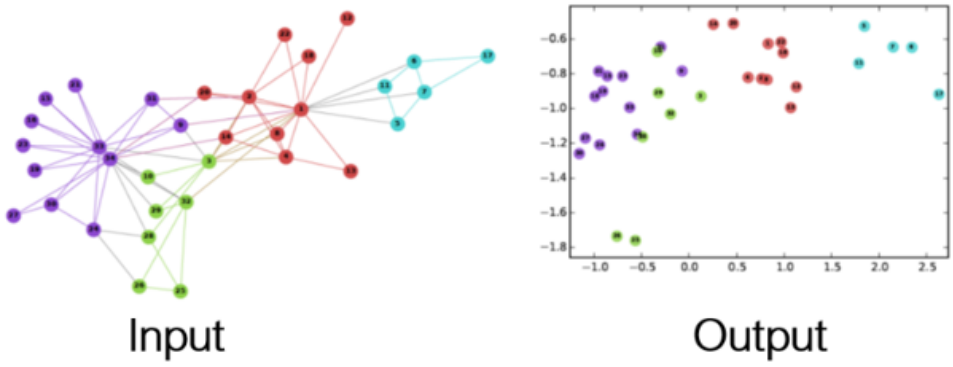
Nokta Çarpımı Benzerliği (*Dot Product Similarity:*) İki vektör arasındaki nokta çarpımı, vektörlerin benzerlik düzeyini belirler. Formülü:

$$\text{Nokta Çarpımı Benzerlik} = \sum_{i=1}^{n} (V_{\text{hastalık}_i} \cdot V_{\text{semptom}_i}) \tag{4.3}$$

Bu yöntemlerle, hastalık ve semptom vektörleri arasındaki mesafeler hesaplanarak, semantik olarak birbirine yakın olan hastalık ve semptomlar belirlenmiştir. Bu benzerlik yöntemleri, farklı perspektiflerden benzerlik ölçümleri sağlayarak, hastalık ve semptom vektörlerini ilişkilendirir. Kosinüs benzerliği, vektörlerin yönlerine odaklanarak, büyüklüklerinden bağımsız olarak semantik yakınlığı değerlendirir. Öklid benzerliği, vektörler arasındaki fiziksel mesafeyi ölçer ve daha somut bir yakınlık duygusu sağlar. Nokta çarpımı benzerliği ise, bir vektörün diğerine ne kadar "uyduğunu veya projekte edildiğini değerlendirerek, elemanların eşleşme derecesine göre bir benzerlik sunar.

4.3 Graf Embedding Yöntemleri ile İlişki Çıkarımı

Graf embedding tabanlı ilişki çıkarımı, karmaşık ağ yapılarındaki ilişkileri ve etkileşimleri anlamak için kullanılan bir yöntemdir. Sosyal ağlardan biyolojik ağlara, bilgi ağlarından iletişim ağlarına kadar birçok alanda büyük ve karmaşık graf yapıları bulunmaktadır. Bu yapılar içerisindeki düğümler arasındaki ilişkiler anlamlı bilgiler içerdiğinden bu bilgiler çıkarılarak çeşitli uygulama alanlarında kullanılabilir. Graf embedding yöntemleri, grafın düğümlerini, kenarlarını ve potansiyel olarak bu öğeler arasındaki daha yüksek düzey ilişkileri düşük boyutlu bir vektör uzayına gömmeyi hedefler. Bu sayede, graf yapıları üzerinde makine öğrenimi ve derin öğrenme tekniklerini uygulamak mümkün hale gelir. İlişki çıkarımı sürecinde, bu gömülü temsiller kullanılarak düğümler arasındaki potansiyel ilişkiler tahmin edilir. Şekil 4.5'te graf embedding yöntemi gösterilmiştir.



Şekil 4.5: Graph Embedding Yöntemi

Biyomedikal bilgi grafları, genler, proteinler, ilaçlar, hastalıklar ve semptomlar gibi biyolojik varlıkları düğümler olarak ve bu varlıklar arasındaki etkileşimleri, ilişkileri veya yolları kenarlar olarak modelleyen yapılardır.

Graph embedding yöntemleri ile büyük bilgi graflarında yer alan düğümler ve kenarlar vektör temsillerine dönüştürülerek yüksek boyutlu veriler işlenebilir hale gelmektedir. Bu temsillerin makine öğrenimi modellerine girdi olarak verilmesi sonucunda hastalık tahmininden ilaç yeniden konumlandırmaya, gen fonksiyonunun çıkarılmasından kişiselleştirilmiş tedavi stratejilerinin geliştirilmesine kadar geniş bir yelpazede biyomedikal sorunlar üzerinde çalışmak mümkündür.

Biyomedikal bilgi grafları, her biri belirli ilişki türüyle bağlı olan iki ilgili varlık ve aralarındaki ilişkiyi ifade eden üçlüler kullanılarak yapılandırılır. Örneğin, COVID-19, karakterize_ edilir, ateş üçlüsünde ilk ve üçüncü kavram aralarında ilişki bulunan medikal varlıkları ortadaki kavram ise bu medikal varlıklar arasındaki ilişki türünü temsil etmektedir. Biyomedikal bilgi grafı oluşturma aşamasında üçlülerin elde edilmesi önemli bir aşamayı oluşturmaktadır. Bu aşamada, çeşitli dilbilimsel kurallar ve uzman bilgisi kullanılarak üçlüler oluşturulmaktadır. Tez kapsamında geleneksel yöntemlere karşı güncel yaklaşımlar arasında yer alan Büyük Dil Modelleri'nin kullanımı, bu üçlülerin otomatik olarak türetilmesi ve zenginleştirilmesi için önerilmektedir. Bu kapsamda kullanılan GPT ve Google LaMDA dil modellerine ilişkin bilgi alt bölüm 4.4'te verilmiştir.

4.4 Büyük Dil Modelleri

4.4.1 GPT Modelleri

OpenAI tarafından geliştirilen GPT modelleri (Radford et al., 2018), dikkat mekanizmalarına dayalı Transformer mimarisini kullanır. Bu mimari, modelin bir cümledeki her kelimenin önemini, diğer tüm kelimelerin bağlamında değerlendirmesine olanak tanır. GPT, verilen bir prompta yanıt üretirken, önceden eğitimli bir Transformer mimarisi kullanır. Transformer mimarisi, dikkat mekanizmaları kullanarak bir dizinin farklı elemanları arasındaki bağlamlı ilişkileri modelleyebilmektedir. RNN ve CNN gibi önceki yaklaşımlara kıyasla, uzun mesafeli bağlantıları daha etkin bir şekilde öğrenebilmektedir. İlişki çıkarımı gibi görevler sırasında metin içerisindeki kelimeler arasındaki bağlamsal ilişkiler tespit edilerek bu bağlamda mantıklı çıkarımlar yapılmaktadır. Transformer mimarisinin temel bileşenleri ve çalışma şekli aşağıda açıklanmıştır.

- Embedding Katmanı: Bu katmanda her kelime veya token, modelin anlayabileceği bir vektöre dönüştürülür. Bu vektörler, kelimenin semantik bilgisini ve kullanım bağlamını içerir.
- Positional Embedding: Transformer mimarisi sıralı bilgiyi doğrudan işleyemediğinden her kelime vektörüne pozisyon belirten kodlamalar eklenir. Bu sayede model, kelimenin cümledeki sırasını da dikkate alır.
- Dikkat Mekanizması Dikkat mekanizması, bir dizideki her elemanın diğer tüm elemanlarla olan ilişkisini ağırlıklandırarak, hangi elemanlara dikkat edilmesi gerektiğini belirler. Bu mekanizmanın formülü şu şekildedir:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.4}$$

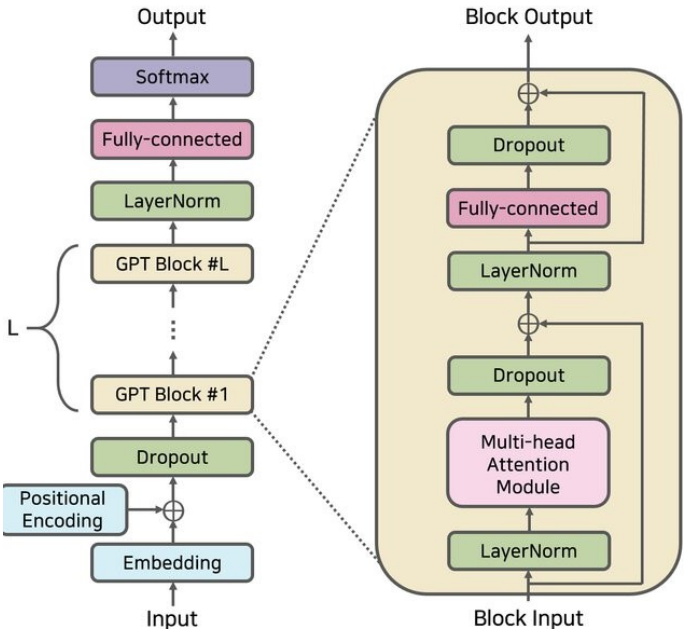
  Burada:

  - Q (sorgu), K (anahtar), V (değer) matrisleridir ve her biri dizinin elemanlarına karşılık gelen temsillerdir.
  - d_k, anahtar vektörlerin boyutudur ve softmax skorlarını düzgünleştirmek için kullanılır.

  - QK^T işlemi, sorgular ile anahtarlar arasındaki uyumluluğu ölçer ve her elemanın diğer elemanlarla olan ilişkisinin ağırlığını hesaplar.
  - Softmax fonksiyonu, bu ağırlıkları olasılık dağılımına dönüştürür, böylece her eleman için diğer elemanlara göre "dikkat" dağılımı elde edilir.

- İleri Beslemeli Ağlar: Dikkat mekanizmasından elde edilen çıktılar, ardışık olarak konumlandırılmış sinir ağlarından geçirilir. Bu ağlar, modelin çıkarımlar yapabilmesi için gerekli karmaşıklığı ve kapasiteyi sağlar.
- Çıktı Katmanı: Modelin son katmanı, kelime vektörlerini geriye doğru kelime dağarcığına çevirir ve en yüksek olasılıklı kelimeyi veya tokeni seçer. Bu süreç, verilen prompta ifade edilen ilişkileri anlama ve bu bağlamda mantıklı yanıtlar üretme kapasitesine katkıda bulunur.

GPT mimarisi Şekil 4.6 ile gösterilmiştir. GPT'nin her adımında gerçekleşen bu işlemler modelin metin içerisindeki ilişkileri çıkarmasını ve bu bilgiye dayanarak doğru yanıtları üretmesini sağlar. İlk GPT modeli 2018'de duyrulmuştur ve 117 milyon parametreye sahiptir. Sonrasında geliştirilen modellerden GPT-2 modeli 1.5 milyar, GPT-3 175 milyar parametreye sahiptir. GPT-4 modeli ise 1.76 trilyon parametre ile en gelişmiş GPT modelidir.



Şekil 4.6: GPT Mimarisi (Lee, 2023)

GPT-3 ve sonraki modeller, az örnek ile öğrenme (few-shot learning) kabiliyetleriyle ön plana çıkmaktadır. Bu özellik ile modele sadece birkaç örnek verildiğinde bir görevi anlaması ve bu görevi yerine getirebilmesi sağlanır.

GPT-3.5 ve GPT-4 arasındaki farklar, esas olarak boyut, performans ve yetenekler açısından görülür. GPT-4, GPT-3.5'a kıyasla daha karmaşık bir modeldir ve daha fazla parametreye sahiptir. Bu artış, genellikle daha ince nuansların daha iyi anlaşılması, daha kapsamlı bilgi ve geliştirilmiş bağlam farkındalığına yol açar. GPT-3.5, 16k giriş ve 4k çıkış bağlam penceresine sahipken, GPT-4, giriş için 128k'ya kadar ve çıkış için 4k bağlam penceresi sunar. Bu, GPT-4'ün çok daha uzun metinleri okuyup hatırlayabilmesi anlamına gelir, örneğin 200+ sayfa metni işleyebilir.

4.4.2 Google LaMDA

Google LaMDA, diyalog odaklı doğal dil işleme konusunda uzmanlaşmış bir dil modelidir. LaMDA, çeşitli diyalog verileri ile geniş kapsamlı bir eğitimden geçirilmiş olup, birçok konu üzerinde açık konuşmalar gerçekleştirebilme yeteneğine sahiptir. LaMDA'ın temel altyapısında transformer tabanlı PaLM

2 (Anil et al., 2023) modeli de bulunmaktadır. PaLM 2'nin ön eğitim veri seti, web belgeleri, kitaplar, kod, matematik ve konuşma verileri gibi çeşitli kaynaklardan oluşan geniş bir seti içerir. PaLM 2'nin 540-milyar parametreye sahip ve çok dilli (100 adet konuşma dili ve 20 adet programlama dili) ve akıl yürütme yeteneklerine sahip bir dil modelidir. Palm2'nin pathways özelliği sayesinde farklı görevler arasında bilgi transferi yaparak geniş bir kelime ve sembol çeşitliliği arasındaki ilişkileri öğrenmesini sağlar. Ayrıca, transformer mimarisi sayesinde girdi metnini daha iyi anlayarak uygun yanıtlar üretmesini sağlar. Az örnek ile öğrenmeye dayanarak, PaLM 2 az örnekle öğrenme (*few shot learning*) kullandığından, sınırlı bir veri setinden içerdiği bilgiler sayesinde niş bağlamlarda doğru yanıt verebilir. PaLM 2'nin akıl yürütme ve çıkarım yapma yetenekleri, birçok veri seti üzerinde yapılan karşılaştırmalarda (Tablo 4.1), bazı durumlarda GPT-4 ile yakın sonuçlar elde ederken, bazı veri setlerinde ise GPT-4'ü aşarak daha üstün performans sergilediği görülmüştür.

Tablo 4.1: Akıl yürütme veri kümelerinde farklı modellerin performans karşılaştırması (Anil et al., 2023).

Dataset	**SOTA**	**GPT-4**	**PaLM**	**PaLM 2**
WinoGrande	87.5%	87.5%	85.1%	90.9%
ARC-C	96.3%	96.3%	88.7%	95.1%
DROP	88.4%	80.9%	70.8%	85.0%
StrategyQA	81.6%	-	81.6%	90.4%
CSQA	-	-	80.7%	90.4%
XCOPA	89.9%	-	89.9%	94.4%
BB Hard	65.2%	-	65.2%	78.1%

Ancak Palm2 eğitim veri setinden çok farklı bir veri ile karşılaştığında uygun yanıt üretmekte zorluk yaşayabilir. Ayrıca, eğitim yapma ve çıkarım sürelerinin uzun olması bu modelin dezavantajları arasında sayılabilir.

4.5 Graf Tamamlama Yöntemleri

Bilgi grafları, belirli alanlarla ilgili geniş bilgi birikimi sunmalarına karşın, sağlık gibi sürekli gelişen alanlarda bazen bilgilerin eksik kalabildiği veya bazı potansiyel bağlantıların henüz keşfedilmediği durumlar ortaya

çıkabilmektedir. Graf tamamlama yöntemleri bilgi graflarındaki eksik veya potansiyel bağlantıların tahmin edilmesinde önemli bir rol oynamaktadır. Graf gömme modelleri ile oluşturulan varlık ve ilişkilerin vektör temsillerinin analiz edilmesi ile bilgi grafına eklenmesi gereken varlık ve ilişkiler hakkında tahmin yapılabilir. Tez kapsamında, oluşturulan hastalık-semptom bilgi grafı üzerinde potansiyel bağlantıların tespit edilmesi için TransE, ComplEx, DistMult ve HolE yöntemleri kullanılmıştır. Bu yöntemler alt bölümlerde açıklanmıştır.

4.5.1 TransE

TransE (Bordes et al., 2013) bilgi grafındaki varlık ve ilişkileri modellemek için çeviri embeddingleri kullanan bir tekniktir. Bu teknikte, ilişkiler, bir varlığın embedding'ini diğer varlığın embedding'ine çevirecek vektörler olarak modellenir. Bilgi graflarını gömme işleminde, düğümler ve ilişkiler için vektör ataması yapılır ve her bir üçlüde, nesne vektörü ile özne vektörünün ilişki vektörü doğrultusunda yapılan çevirinin mesafesi en aza indirgenerek işlem gerçekleştirilir.

TransE, bir bilgi grafiğindeki her varlık ve ilişki için bir vektör atar. Bir varlık çifti (örneğin, "bronşit" ve "öksürük") ve aralarındaki bir ilişki (örneğin, "neden olur") verildiğinde, TransE bu ilişkiyi vektör uzayında bir çeviri olarak modellemeye çalışır. Yani, "bronşit" vektörüne "neden olur" ilişkisinin vektörü eklenirse, bu toplamın "öksürük" vektörüne yakın olması beklenir.

Verilen bir üçlü (h, r, t) için, h başlangıç, t bitiş ve r ilişki olmak üzere, TransE ilişkiyi bir çeviri olarak modellemektedir: $h + r \approx t$. Geçerli bir üçlü (h, r, t), için skorlama fonksiyonu şu şekildedir:

$$f(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_p \tag{4.5}$$

burada $\mathbf{h}$, $\mathbf{r}$ ve $\mathbf{t}$ sırasıyla başlangıç, ilişki ve bitişin vektör temsillerini ve $\|\cdot\|_p$ L_p normunu belirtir. TransE modeli many-to-many ve asimetrik ilişkileri modellemekte zorlanabilir.

4.5.2 ComplEx

ComplEx (Trouillon et al., 2016), bilgi graflarındaki varlık ve ilişkileri karmaşık sayılar kullanarak temsil eden bir gömme yöntemidir. Bu model, ilişkilerin simetrik olmayan doğasını (yani bir ilişkinin tersinin her zaman geçerli olmaması gibi özellikler) daha etkin bir şekilde yakalayabilmek için geliştirilmiştir. ComplEx, özellikle many-to-many ve asimetrik ilişkiler gibi karmaşık ilişki türlerini modellemek için kullanılır. Bu yöntemin çalışma şekli şöyledir:

- Her bir varlık ve ilişki için karmaşık (yani hem gerçek hem de sanal sayı bileşenleri içeren) vektörler oluşturulur. Bu vektörler, bir bilgi grafiğindeki ilişkileri temsil etmek için kullanılır.
- Her bir üçlünün karmaşık vektör temsillerinin bir iç çarpım işlemi ile üçlülerin geçerliliğini ölçen bir skor üretilir.
- Gerçekleştirilen iç çarpım işlemi sonucunda elde edilen karmaşık sayının gerçek kısmı, üçlünün geçerlilik skorunu verir.
- Her bir varlık ve ilişki için bir karmaşık vektör atandığında, bu vektörlerin boyutları birbiriyle uyumlu olacak şekilde çarpılır.
- Her varlık ve ilişkinin vektör temsili karmaşık sayılardan oluştuğu için, bu çarpım işlemi sırasında hem çarpma hem de karmaşık sayıların eşlenik alınma özelliği kullanılır.

Özetle, ComplEx modelinde ilişkiler, karmaşık sayılar kullanılarak temsil edilir ve bu temsillerin iç çarpımı alınarak üçlünün model tarafından ne kadar makul bulunduğu bir skorla ifade edilir. ComplEx'in skorlama fonksiyonu aşağıda yer verilmiştir. Burada h, r, ve t sırasıyla bir üçlünün öznesi, ilişkisi ve nesnesinin karmaşık vektör temsillerini, k bu vektörlerin her bir bileşeni için bir indeksi ifade eder.

$$f(h, r, t) = \mathrm{Re}\left(\sum_{k=1}^{K} h_k \cdot r_k \cdot \overline{t_k}\right)$$

Burada Re() işlemi bir karmaşık sayının gerçek kısmını alır ve $\overline{t_k}$ işlemi t'nin karmaşık eşleniğini alır. K, gömme vektörlerinin boyutunu ifade eder.

4.5.3 DistMult

DistMult (Yang et al., 2014), temel olarak, varlık ve ilişkileri düşük boyutlu vektörler olarak temsil ederek, bu temsillerin çarpımı yoluyla varlık çiftleri arasındaki olası ilişkilerin yoğunluğunu veya olasılığını değerlendirir. DistMult modelinde her varlık ve ilişki, bir vektör ile temsil edilir. Bir varlık çifti arasındaki bir ilişkinin var olma olasılığı, bu varlıkların ve ilişkinin vektör temsillerinin skaler çarpımı ile hesaplanır. Bu işlem, varlıklar ve ilişkiler arasındaki ilişkinin yoğunluğunu veya gücünü nicelendirmeye yardımcı olur. DistMult metodunun temel formülü şu şekilde ifade edilir:

$$f(h, r, t) = \mathbf{e}_h^T \mathbf{r}_r \mathbf{e}_t \tag{4.6}$$

Burada $\mathbf{e}_h$ ve $\mathbf{e}_t$ başlangıç ve bitiş varlıklarının gömü vektörleri, $\mathbf{r}_r$ ilişkinin gömü vektörüdür. DistMult, basit ve etkin bir yöntem olmasına rağmen asimetrik ilişkileri modelleme konusunda kısıtlamalara sahiptir.

4.5.4 HolE

HolE (Nickel et al., 2016), Holographic Embeddings bilgi graflarındaki varlık ve ilişkileri gömmek için dairesel korelasyonu kullanarak karmaşık etkileşimleri modelleyen bir yöntemdir. HolE, bilgi graflarındaki asimetrik ve many-to-many ilişki türlerini daha iyi yakalamak için tasarlanmıştır. HolE, her varlık için düşük boyutlu bir vektör atar ve bu vektörler arasındaki ilişkileri dairesel korelasyon kullanarak modeller. Dairesel korelasyon, iki vektörün öğelerini kaydırarak çarpıp toplamını alarak yeni bir vektör üretir. BU sayede, ilişkiler zengin ve çok yönlü bir şekilde ifade edilir.

Bir h (başlangıç varlık), r (ilişki) ve t (bitiş varlık) üçlüsü için HolE modelinin skorlama fonksiyonu şu şekilde ifade edilir:

$$f(h, r, t) = \mathbf{r}^\top (\mathbf{e}_h \star \mathbf{e}_t)$$

Burada $\mathbf{e}_h$ ve $\mathbf{e}_t$ başlangıç ve bitiş varlıklarının vektör temsilleridir; $\mathbf{r}$ ise ilişkinin vektör temsilidir. $\star$ işlemi, $\mathbf{e}_h$ ve $\mathbf{e}_t$ vektörleri arasındaki dairesel korelasyonu ifade eder ve bu, iki vektörün elemanlarının bir tür iç içe geçirilmiş çarpımını temsil eder.

Graf tamamlama yöntemlerinin karşılaştırılması Tablo 4.2'de verilmiştir.

Tablo 4.2: Graf Tamamlama Yöntemlerinin Karşılaştırması

Model	**Ana Fikir**	**Avantajlar**	**Sınırlılıklar**
TransE	İlişkileri çeviri olarak modelleme	Basit ve verimli, hızlı eğitim ve çıkarım süreçleri	Karmaşık ve çok-çok ilişkileri modellemede zorlanır
ComplEx	Karmaşık sayılar kullanarak gömme	Asimetrik ve çok-çok ilişkileri etkili bir şekilde modelleyebilir	Karmaşık hesaplama yapısı
DistMult	Diyagonal matris kullanarak ilişkileri modelleme	Basit yapı, simetrik ilişkileri iyi modelleme	Asimetrik ilişkilerde sınırlı
HolE	Dairesel korelasyon kullanarak gömme	Karmaşık ilişkileri ve etkileşimleri modelleyebilir	Yüksek hesaplama maliyeti

4.6 Yeni Bağlantıların Tahmin Edilmesi

Model eğitimleri sonucunda en başarılı yöntem olan TransE kullanılarak potansiyel bağlantılar tespit edilmiştir. Bu kapsamda, Random Uniform (Rastgele Örnekleme), Graph Degree (graf derecesi), Entity Frequency (varlık sıklığı), Cluster Coefficient (Küme Katsayısı), Cluster Triangles (Küme Üçgenleri) ve cluster squares (Küme kareleri) yaklaşımları kullanılmıştır. Bu yaklaşımlar çeşitli kriterlere göre örneklemeler yaparak veri seti içerisindeki üçlülerden aday setleri oluşturulabilir. Bu yöntemler aşağıda açıklanmıştır.

Random Uniform (Rastgele Örnekleme): Bu yöntem, veri seti veya ağdaki tüm varlıklar arasından eşit olasılıkla rastgele seçimler yapar. Her bir varlığın seçilme şansı eşittir.

Graph Degree (Graf Derecesi): Bir ağdaki düğümlerin derecesine (bağlantı sayısına) göre ağırlıklı örnekleme yapar. Yüksek dereceye sahip düğümler, daha az bağlantıya sahip olanlara göre daha sık seçilir.

Entity Frequency (Varlık Sıklığı): Varlıkların frekanslarına göre ağırlıklı bir örnekleme yapılır. Daha sık görünen varlıklar daha yüksek olasılıkla seçilir.

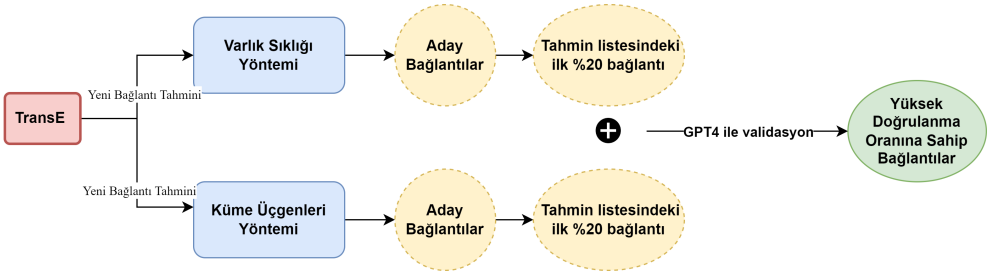
Cluster Coefficient (Küme Katsayısı): Düğümlerin kümeleme katsayısına

göre ağırlıklı örnekleme yapılır. Kümeleme katsayısı, bir düğümün komşuları arasındaki bağlantı yoğunluğunu ölçer. Yüksek kümeleme katsayısına sahip düğümler, daha yüksek bir olasılıkla seçilir.

Cluster Triangles (Küme Üçgenleri): Düğümlerin oluşturduğu üçgen sayısına göre ağırlıklı örnekleme yapar. Üçgenler, bir ağdaki üç düğüm arasında karşılıklı bağlantıları ifade eder. Bu yöntem, daha sıkı bağlantılı düğüm gruplarını örneklemede kullanılır.

Cluster squares (Küme kareleri): Düğümlerin oluşturduğu kare (dört düğüm arasındaki döngüsel bağlantılar) sayısına göre ağırlıklı örnekleme yapar.

Yukarıda açıklanan yöntemler ile tespit edilen farklı bağlantılar GPT4 dil modeli aracılığı ile bilimsel literatürdeki bilgiler kullanılarak otomatik doğrulanmıştır. GPT-4 tarafından üretilen sonuçlar insan doğrulamasından geçirilmiştir. Daha doğru bağlantılar ürettiği gözlenen varlık sıklığı ve küme üçgenleri yöntemleri hibrit kullanılarak elde edilen aday bağlantıların seçimi optimize edilmiştir. Her bir yöntem tarafından tespit edilen bağlantılar, tahmin skorlarına göre sıralanmış ve bu sıralamaların ilk %20'sinde yer alan bağlantılar birleştirilerek daha doğru bağlantılardan oluşan bir aday bağlantı havuzu oluşturulmuştur. Bu yöntem Şekilde 4.7'de gösterilmiştir.



Şekil 4.7: Aday Bağlantıların Seçimi

4.7 Değerlendirme Ölçütleri

Bu bölümde benzerlik yöntemlerinin ve graf tamamlama yöntemlerinin değerlendirilmesine ilişkin ölçütler sunulmuştur.

4.7.1 Benzerlik Yöntemlerinin Değerlendirilmesi

Tez kapsamında ele alınan benzerlik yöntemlerinin hastalık-semptom ilişkilerini tespit etmedeki başarılarını değerlendirmek amacıyla bir benzerlik skoru hesaplanmıştır. Değerlendirme işleminde, hastalığa ait bilinen semptomların üst sıralarda tespit edilme oranı esas alınmıştır. Benz skoru, doğrulanmış semptomların ilk 50, 100 ve 200 sıralamasında görülme oranlarının toplamını, ilgili hastalığa ait doğrulanmış semptom sayısına bölerek hesaplanır. Aşağıdaki formül ile hesaplanır:

$$\text{benz\_skoru} = \frac{(\#\text{semptom50} + \#\text{semptom100} + \#\text{semptom200})}{(\#\text{BilinenSemptomlar})} \tag{4.7}$$

Burada:

- #BilinenSemptomlar: İlgili hastalığın Dünya Sağlık Örgütü (WHO) ve Mayo Clinic gibi sağlık kaynakları tarafından doğrulanmış semptom sayısıdır.
- #semptom50: Doğrulanmış semptomların ilk 50 sıradaki görülme sayısıdır.
- #semptom100: Doğrulanmış semptomların ilk 100 sıradaki görülme sayısıdır.
- #semptom200: Doğrulanmış semptomların ilk 200 sıradaki görülme sayısıdır.

Elde edilen değer, benzerlik yönteminin hastalık ve semptomlar arasındaki ilişkiyi ne kadar iyi tahmin ettiğini gösterir. Daha yüksek bir benz skoru, benzerlik yönteminin hastalıkları doğru bir şekilde tespit etme başarısını ifade eder.

Benzerlik yöntemleri kullanılarak elde edilen semptomlar, benzerlik mesafelerine göre sıralandıktan sonra, en az benzer olan 10 semptom üzerinde bilinen veri tabanları ve bilimsel literatürden yararlanarak kapsamlı bir literatür taraması gerçekleştirilmiştir. Bu tarama, bu semptomların hastalıklarla olan ilişkilerinin doğruluğunu ve klinik önemini değerlendirmek için yapılmıştır.

4.7.2 Graf Tamamlama Yöntemlerinin Değerlendirilmesi

Bu bölümde, tez kapsamında kullanılan TransE, DistMult, ComplEx ve HolE graf tamamlama yöntemleri için üç değerlendirme metriği olan MR, MRR ve Hit@Rate açıklanmıştır.

MR, bilgi erişimi veya sorgulamalara yanıt olarak cevapları sıralama gibi görevlerde bir modelin performansını değerlendirmek için kullanılan istatistiksel bir ölçüttür. MR, bir dizi sorgu için doğru veya alakalı ilk cevabın sıralamalarının ortalaması alınarak hesaplanır. MR için formül şöyledir:

$$MR = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \text{rank}_i \tag{4.8}$$

Burada $|\mathcal{Q}|$ sorgu sayısını ve rank_i i-inci sorgu için doğru veya ilgili öğenin sırasını ifade eder.

MRR, bir modelin farklı sorgulara sıralı yanıtlar üretme etkinliğini değerlendirmek için kullanılan bir istatistiksel metriktir. MRR, yanıt listesindeki ilk doğru veya alakalı cevabın ters sıralarının ortalaması alınarak belirlenir. MRR için formül şöyledir:

$$MRR = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \frac{1}{\text{rank}_i} \tag{4.9}$$

Burada $|\mathcal{Q}|$ sorgu sayısını ve rank_i i-inci sorgu için ilk ilgili cevabın konumunu gösterir.

Hit@Rate, tavsiye sistemlerinin doğruluğunu değerlendirmek için kullanılan bir ölçüttür. Kullanıcı için alakalı veya faydalı olan bir önerilen öğe veya öğeler dizisi durumlarının oranı olarak tanımlanır. Hit@Rate için formül şöyledir:

$$\text{Hit@K} = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} 1(\text{rank}_i \leq K) \tag{4.10}$$

Burada $|\mathcal{Q}|$ sorgu sayısını, rank_i i-inci sorgu için ilk ilgili öğenin sıra konumunu ifade eder ve 1 gösterge fonksiyonu, $\text{rank}_i \leq K$ ise 1 değerini, aksi halde 0 değerini döndürür.

5 DENEYSEL ÇALIŞMALAR

Bu bölümde tez kapsamında yapılan deneysel çalışmalara yer verilmiştir. Tüm deneysel çalışmalar Python dili ve ilgili kütüphaneler kullanılarak Anaconda Navigator aracı ile gerçekleştirilmiştir. Bölüm 5.1'de tez çalışması için toplanan özgün veriseti açıklanmıştır. Bölüm 5.2'de benzerlik tabanlı, Bölüm 5.3'te graf tabanlı ilişki çıkarımı için gerçekleştirilen deneysel çalışmalar ile ilgili bilgiler verilmiştir.

5.1 Veri Seti

Tez kapsamında bronşit, COVID-19, astım ve pulmoner emboli hastalıkları ile ilgili semptom ve hastalık isimlerini içeren özetlerden oluşan bir veri seti oluşturulmuştur. Seçilen hastalıklar, solunum sistemi ve akciğerlerde görülen, toplum sağlığını önemli ölçüde etkileyen ve geniş bir popülasyonu kapsayan hastalıklardır. Ayrıca, solunum yolu hastalıkları genellikle benzer semptomlara sahip olduğundan bu durum tanı ve tedavi süreçlerini zorlaştırmaktadır. Bu çalışma kapsamında seçilen bronşit, astım, pulmoner emboli ve COVID-19 solunum yolu hastalıklarının geniş bir yelpazesini temsil etmektedir. Bu hastalıkların belirtileri bazen birbirine benzeyebileceğinden erken teşhis ve etkili tedavi için bu hastalıkların en çok ilişkili olan semptomların belirlenmesi, hastalıkların ayırt edilebilmesi açısından önemlidir. Biyomedikal literatürde solunum yolu hastalıklarının semptomlarını ayırt etmeye yönelik klinik araştırmalar mevcuttur (Ma et al., 2018; Van der Sar et al., 2023) Bronşit, COVID-19, astım ve pulmoner emboli hastalıkları ile ilgili makale özetlerini elde etmek için tıp alanı ile ilgili çok sayıda makale, kitap ve olgu sunumunun yer aldığı PubMed veri tabanından faydalanılmıştır. BioPython kütüphanesi kullanılarak web madenciliği ile çalışma kapsamında kullanılan bilimsel makale özetleri otomatik olarak elde edilmiştir. Özetlerin hastalık ve semptom adlarını içermesini sağlamak için, belirli sorgu cümleleri kullanılmıştır. Sorgu cümleleri oluşturulurken çalışmaya konu olan dört hastalık ve bu hastalıkların sağlık kaynaklarında belirtilen semptomları kullanılmıştır.

Ayrıca, daha çok makale özetine ulaşmak için semptom ve hastalıkların eş anlamlıları da sorgulara dahil edilmiştir. Bu yöntemle, farklı hastalık ve semptom isimlerini içeren makale özetlerinin elde edilmesi amaçlanmıştır. Sorgu işleme ve veri toplama için BioPython kütüphanesinin BioEntrez yöntemi kullanılmıştır. Sonuç olarak, 120 sorgu cümlesi ile COVID-19, bronşit, astım ve pulmoner emboli hastalıklarına ait sırasıyla 5718, 4320, 1948 ve 4208 adet özet olmak üzere toplamda 16.194 makale özeti elde edilmiştir. Verisetine ait bilgiler Tablo 5.1'de gösterilmiştir. Örnek sorgu cümleleri şu şekildedir:
coronavirus disease[Title/Abstract] AND loss of taste[Title/Abstract]
bronchitis[Title/Abstract] AND sore throat [Title/Abstract]
asthma[Title/Abstract] AND cough[Title/Abstract]
pulmonary embolism[Title/Abstract] AND chest pain [Title/Abstract]

Bu sorgular yardımıyla, hastalık ve semptom adlarını içeren makale özetleri elde edilmiştir. Böylece solunum yolu hastalıkları ile ilgili ilişki çıkarımına yönelik bir veri seti oluşturulmuştur. Bildiğimiz kadarıyla bu kapsamda oluşturulan ilk veri setidir. Veri setine `https://github.com/azerceliktenn/abstracts_for_diseases` linkinden erişim sağlanabilir.

5.2 Benzerlik Tabanlı İlişki Çıkarımı

Hastalık ve Semptom İsimlerinin Elde Edilmesi: Pubmed makale özetlerinde geçen hastalıklar ve semptomların otomatik olarak çıkarılması için Bölüm 4.1'de önerilen medikal NER yöntemi kullanılmıştır. Medikal NER işlemi, Scispacy kütüphanesinin özellikle biyomedikal metinler için

Tablo 5.1: Veri Setine Ait Bilgiler

Solunum Yolu Hastalığı	**Özet Sayısı**
COVID-19	5718
Bronşit	4320
Astım	1948
Pulmoner Emboli	4208
Toplam	**16.194**

optimize edilmiş en_ner_bc5cdr_md dil modeli kullanılarak gerçekleştirilmiştir. Bu model, geniş bir biyomedikal literatür veritabanından öğrenilen derin öğrenme teknikleri sayesinde yüksek doğrulukta hastalık ve kimyasal maddeleri tanıyabilmektedir. Ayrıca, semptomların tanımlanması ve sınıflandırılması için semptom ontolojisi entegre edilmiştir. Bu ontoloji, semptomların daha geniş bir bağlamda anlaşılmasını ve ilişkili medikal durumlarla olan bağlantılarının kurulmasını sağlar. Bu işlem her bir Pubmed makale özet için uygulanarak hastalık ve semptomların çıkarımı sağlanmıştır. Makale özetlerinden semptom ve hastalık çıkarıma dair örnek Şekil 5.1'de verilmiştir.

Şekil 5.1: Makale özetlerinden hastalık ve semptomların çıkarımı

UMLS Normalizasyonu: UMLS, tıbbi ve sağlık alanlarındaki farklı terminolojileri ve kodlamaları bir araya getirmek ve birleştirmek amacıyla geliştirilmiş bir sistemdir (Aronson, 2001). UMLS normalizasyonu, farklı terminolojilerde ve kodlamalarda bulunan tıbbi terimleri veya kavramları, UMLS tarafından tanımlanan ortak kavramlarla eşleştirme ve bu kavramları standart bir formatta temsil etme sürecidir. Bu normalizasyon süreci, farklı sağlık kuruluşları, tıbbi araştırmacılar ve yazılım geliştiricileri için farklı terminoloji sistemleri arasında veri paylaşımını ve iş birliğini kolaylaştırmayı amaçlar. Bu normalizasyon ile farklı terminolojilerin ve kodlamaların karmaşıklığını azaltılır, sağlık verilerinin tutarlılığını ve anlamını arttırır ve tıbbi bilgi paylaşımını ve analizini kolaylaşır. Varlık bağlama, metindeki varlıkların bir bilgi tabanında karşılık gelen terimlerle otomatik olarak eşleştiren bir süreçtir

(Wada et al., 2018). Bu çalışmada bilgi tabanı olarak UMLS kullanılmıştır. UMLS'te yer alan CUI tanımlayıcıları, tıbbi literatürde ve sağlık alanlarında benzer terimleri benzersiz bir şekilde tanımlamak amacıyla kullanılan alfa-numerik kimliklerdir. CUI'ler, tıbbi kavramların farklı adlandırmalarına karşılık gelen anlamlarını ve ilişkilerini belirlemek için kullanılırlar. Örneğin, boğaz ağrısı anlamına gelen sore throat, throat pain, pharyngalgia ifadelerinin tümü UMLS veri tabanında tek bir CUI (C0242429) ile ifade edilmektedir. UMLS normalizasyonu için scispacy tarafından sunulan EntityLinker (varlık bağlayıcı) kütüphanesi kullanılmıştır. Varlık bağlama işlemi elde edilen tüm semptomlara uygulanarak semptom isimlerinin farklı kullanım şekilleri ve eş anlamlılık nedeniyle ortaya çıkabilecek tutarsızlıklar önlenmiş olur. Örneğin Şekil 5.1'de chills ve rigors titreme anlamına gelen semptomlardır. Bu isimlerin tek bir CUI ile ifade edilmesi anlam bütünlüğünün sağlanması açısından önemlidir.

Bu kapsamda elde edilen Tıbbi metinlerde farklı şekillerde yer alan semptomlar ve CUI karşılıklarına ait örnekler Tablo 5.2'de yer almaktadır. Elde edilen tüm semptomların ve hastalıkların UMLS CUI eşleştirmeleri gerçekleştirilmiştir.

Tablo 5.2: Makalelerdeki Farklı Semptom İfadeleri ve CUI Eşleştirmeleri

Semptom Adlandırmaları	**Türkçe Karşılığı**	**UMLS CUI**
sore throat, pharyngalgia, throat pain, pain in throat	boğaz ağrısı	C0242429
headache, cephalgia, head pains	baş ağrısı	C0018681
skin redness, skin erythema, erythemas	cilt kızarıklığı	C0041834
coughing, dry cough, coughs	öksürük	C0010200
pruritus, itching, itch of skin	kaşıntı	C0033774

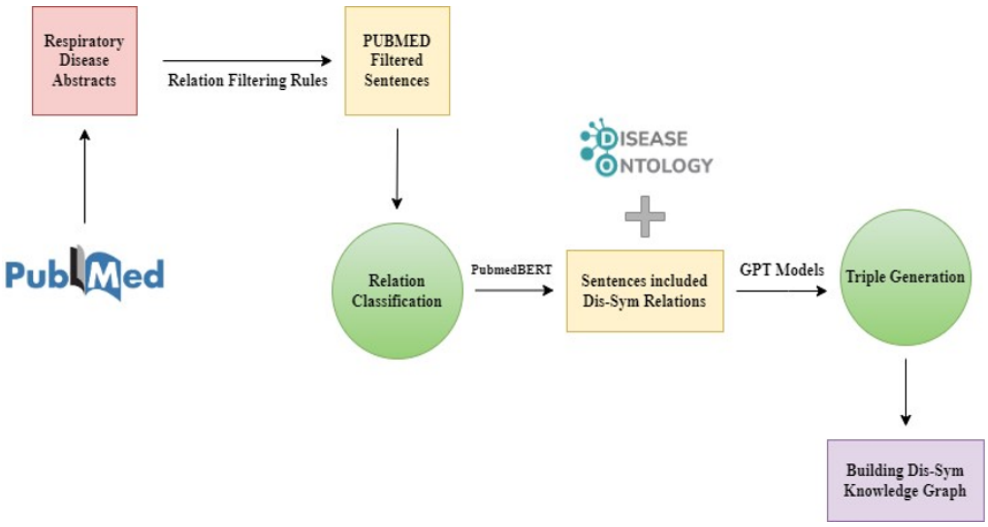
Benzerlik Yöntemlerinin Uygulanması: Hastalıklar ve semptomlar arasındaki semantik ilişkileri tespit edebilmek amacıyla her bir hastalık vektörünün o hastalığa ait özetlerden elde edilen semptomların vektörleri ile benzerlik hesaplaması gerçekleştirilmiştir. İlgili vektörler Scispacy NER modelinden elde edilmiştir. Bu model, BC5CDR veri seti üzerinde eğitilmiştir. Bu vektörler, geniş bir biyomedikal literatür korpusundan eğitildiğinden biyomedikal terimlerin karmaşık bağlamlarını ve kullanımlarını anlama açısından

başarılıdır. Bu kapsamda hastalık ve semptom vektörleri arasındaki, kosinüs, öklid ve nokta çarpımı uzaklıklar hesaplanmıştır. Kosinüs benzerliği, vektörler arasındaki açısal mesafeyi ölçerken, Öklid uzaklığı doğrudan mesafeyi hesaplar ve nokta çarpımı ise vektörlerin doğrudan çarpım sonucu elde edilen skorları kullanır.

5.3 Graf Tabanlı İlişki Çıkarımı

Bu bölümde hastalıklar ve semptomlar arasındaki ilişki analizinde graf tabanlı yöntemler önerilmiştir. İlişki analizi için bir bilgi grafı oluşturulmuş ve bu graf üzerinde graf embedding yöntemleri ile analizler gerçekleştirilmiştir. Yöntem detaylarına alt başlıklarda yer verilmiştir.

Hastalık-Semptom Bilgi Grafının (DS-KG) Oluşturulması: Bu tezde bilimsel literatürden elde edilen medikal makale özetleri kullanılarak bir hastalık-semptom bilgi grafı oluşturulmuştur. Sırasıyla, bilgi grafı için özetlerden uygun cümleleri seçebilmek amacıyla kural tabanlı ve cümle sınıflandırma modelleri geliştirilmiş, ardından büyük dil modelleri kullanılarak cümlelerdeki ilişkiler analiz edilerek hastalık, ilişki, semptom üçlüleri elde edilmiştir. İnşa edilen bilgi grafı üzerinde graf tamamlama, kümeleme ve potansiyel bağlantıların keşfedilmesi için graf embedding tabanlı yöntemler uygulanmıştır. Hastalık-Semptom Bilgi Grafının oluşturulmasına ilişkin yöntem Şekil 5.2'de gösterilmiştir.



Şekil 5.2: Hastalık-Semptom Bilgi Grafının Oluşturulması

Kural Tabanlı Cümle Seçimi: Solunum yolu hastalıklarına ait Pubmed makale özetlerinden hastalıklar ile semptomlar arasındaki ilişkileri tanımlayan cümleleri tespit etmek için kurallar belirlenmiştir: Kural 1: Her cümle en az iki varlık içermelidir; bunlar hastalık veya semptom olabilir. Kural 2: Hastalık-semptom veya hastalık-hastalık çiftleri arasındaki ilişkileri belirten "characterized by", "associated with", "caused by" veya "causes" gibi belirli dil bağlaçlarının bulunması zorunludur.

Dört solunum hastalığı için bu kuralları içeren tekrarlayan cümleler filtrelendikten sonra, toplamda 4023 adet benzersiz cümle belirlenmiştir. Bu cümlelere ek olarak Disease ontolojideki (Disease Ontology Consortium, n.d.) hastalıkların tanımlarında geçen cümlelerden, "characterized by" bağlacı ile "pulmoner" ve "lung" ifadelerini içeren 227 adet cümle aday cümle havuzuna eklenmiştir. Elde edilen cümleler potansiyel olarak hastalık ve semptomlar arasında bir ilişki içerebilir ancak, bazı cümleler bu iki kuralı karşılasa da, her zaman semantik bir ilişki türetmek mümkün olmayabilir. Kural bazlı filtreleme sürecinden sonra elde edilen örnek cümleler 5.3' da gösterilmiştir.

Cümle Sınıflandırma Modellerinin Geliştirilmesi: Kural tabanlı filtreleme ile elde edilen cümlelerin her zaman hastalık ve semptom arasında semantik bir ilişki göstermemesi sorununu çözmek için transformer tabanlı yöntemler kullanılarak cümleler sınıflandırılmıştır. Bu şekilde, belirli bir

Tablo 5.3: Kural bazlı filtreleme sürecinden sonra elde edilen örnek cümleler

Örnek İlişki Cümleleri	**Hastalık-Semptom İlişkisi Varlığı**
Obesity was also significantly associated with bronchitis (adjusted odds ratio (AOR) and 95% confidence intervals (95%CI): 5.29 (2.58;10.85) and with the use of antibiotics (AOR (95%CI): 1.79 (1.09;2.93)). (Pubmed Id:20380692) (Pubmed Id:20380692)	Evet
Children are assessed at ages 9-11, but it is associated with moderate increases in diagnoses of anxiety or depression-which are concentrated among children living in market-rate housing. (Pubmed Id: 31479371)	Hayır

cümlenin hastalık ve semptom arasında bir ilişki içerip içermediğini tahmin eden modeller oluşturulmuştur. İkili sınıflandırma modelini eğitmek için, kural tabanlı filtrelemeden 1535 cümle semantik açıdan incelenmiştir. Bu cümlelerden 1100'ü hastalık ve semptom arasında bir ilişki içerdiği tespit edilerek pozitif olarak, İlişki içermeyen kalan 435 cümle negatif olarak etiketlenmiştir. Sınıflandırma modellerini eğitmek için, BERT mimarisini kullanan dil modellerinden yararlanılmıştır. Bu modeller, kendinden dikkat mekanizmalarına dayanan Transformer mimarisini kullanır. Birkaç katmanlı çift yönlü transformer kodlayıcılardan oluşur. Her model, dil yapıları ve bağlamlarını anlayabilmesi için geniş genel veya özelleşmiş bir derlem üzerinde önceden eğitilir. Önceden eğitilmiş model, daha küçük, göreve özgü bir veri seti kullanılarak ince ayar yapılır. Fine-tuning işleminde, sinir ağı ağırlıkları belirli sınıflandırma görevine göre ayarlanır. Sınıflandırma için fine-tuning yapılırken bir sınıflandırma katmanı modele eklenir. Bu katman, transformer kodlayıcılarının çıktısına dayanarak nihai tahmini oluşturur. Verilen giriş metni için model, sınıf olasılıkları çıkarır ve en yüksek olasılığa sahip sınıf tahmin olarak seçilir. Sınıflandırma için, BERT giriş olarak belirteç dizilerini alır. İlk belirteç benzersiz bir sınıflandırma belirteci [CLS]'dir. Bu belirteçle ilişkilendirilen son gizli durum, sınıflandırma amaçları için dizinin kolektif temsili olarak hizmet eder. Basit bir sınıflandırma katmanı, [CLS] belirteci için transformer çıktısının üstüne eklenir. Bu katman belirli bir görev (örn. duygu analizi, metin sınıflandırması) için eğitilir. İlişki sınıflandırma problemi için kullanılan transformer tabanlı modellerin mimarisi Şekil 5.3'de gösterilmiştir.

Transformer tabanlı modeller arasında, BERT-Base-çokdilli, BioBERT, SciBERT ve PubMedBERT modelleri kullanılmıştır. Bu modeller, genel ve alan özgü derlemler üzerindeki geniş öncül eğitimlerini kullanarak dilin semantiklerini ve yapısını anlamakta ve bu karmaşık görev için son derece etkili olmaktadır. Cümle düzeyi ilişkileri sınıflandırmak için kullanılan transformer modellerini kısaca açıklanmıştır.

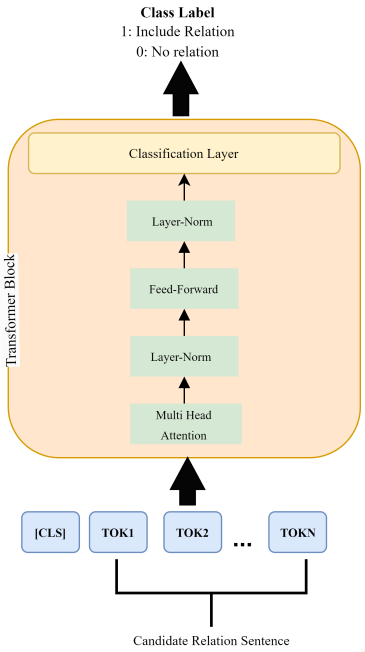
BERT (Devlin et al., 2018), NLP alanında transformer ve dikkat mekanizmalarını içeren bir yaklaşımdır. Tasarımı, her katmanda sol ve sağ bağlamları eş zamanlı olarak dikkate alan derin çift yönlü temsillerin ön eğitimine odaklanır. Sonuç olarak, bu ön eğitimli BERT modeli, çeşitli görevler için uyum sağlanabilir, sadece ince ayar için tek bir ek çıkış katmanı gerektirir ve göreve özgü mimaride önemli değişiklikler yapılmasına gerek kalmaz.

Sci-BERT (Beltagy et al., 2019), büyük bir bilimsel metin koleksiyonu üzerinde özel olarak eğitilmiş BERT modelinin bir uyarlamasıdır. Bilim alanına BERT'in genişletilmesi, geniş, disiplinler arası bir bilimsel metinler derlemi üzerinde ön eğitim yapmayı içerir. Bu yaklaşım, sonraki bilimsel NLP görevleri için transfer öğrenim verimliliğini artırır.

BioBERT (Lee et al., 2019), geniş biyomedikal derlemler üzerinde önceden eğitilmiş özelleşmiş bir dil temsil modelidir. Amacı, BERT modelini biyomedikal metin madenciliği uygulamaları için rafine etmektir. BioBERT, önceki modellere göre önemli ilerlemeler göstermiş ve çeşitli biyomedikal metin madenciliği görevlerinde, örneğin biyomedikal adlandırılmış varlık tanıma, ilişki çıkarma ve soru cevaplama gibi alanlarda başarılı olmuştur.

PubMedBERT (Gu et al., 2021), yalnızca PubMed veritabanının özetleri ile eğitilmiş olup, büyük miktarda biyomedikal literatür içerir. Bu model, biyomedikal araştırmalarda kullanılan dili daha iyi anlamak ve işlemek için ince ayar yapılmıştır. Bu, genel derlemler üzerinde ön eğitim yapılmış modellere kıyasla çeşitli biyomedikal NLP görevlerinde performansı iyileştirir.

Kurallarla belirlenen tüm cümleler, diğer modellere göre daha üstün performans sergileyen PubMedBERT sınıflandırma modeli tarafından tekrar tahmin edilir. Eğer cümle pozitif olarak tahmin edilirse, yani cümle hastalık



Şekil 5.3: Fine tuning of transformer-based models for relation classification

ve semptom arasında bir ilişki içeriyorsa, bu cümle graf veri setini oluşturmak için kullanılacak olan nihai cümle havuzuna eklenmiştir. Tüm özetler kural ve sınıflandırma tabanlı filtrelerden geçirildikten sonra graf veri seti için 3191 adet cümle elde edilmiştir.

Büyük Dil Modelleri ile Üçlü Üretimi: Solunum yolu hastalık ilişkilerini içeren bir bilgi grafı inşa etmek için grafın düğüm ve kenarlarını temsil edecek üçlülere (subject, predicate, object) ihtiyaç bulunmaktadır. (Örneğin, covid-19, ilişkilidir, nefes darlığı) Bu kapsamda, cümlelerden hastalık-semptom ilişkilerinin üçlü olarak çıkarılması için büyük dil modellerinin çıkarım özelliğinden faydalanılmıştır. GPT ve LaMDA dil modellerinin sağladıkları API aracılığı ile üçlüler elde edilmiştir. Kural tabanlı ve sınıflandırma modellerinin birlikte kullanılarak oluşturulan cümle havuzu (3191 adet cümle), dil modeline girdi olarak verilmiş ve üretilen üçlüler çıktı olarak geri döndürülmüştür. Kullanıcı isteklerinden üçlüler üretmek için GPT tabanlı modellerin API'si ve Google LaMDA'ın kullanıcı arayüzü (Google Bard) kullanılmıştır. Üçlü üretimi için aşağıdaki komut kullanılmıştır.

"Can you extract respiratory diseases-symptoms relation triples from

the given sentences that consist of relation types as associated with, caused by, causes, and characterized by for training the graph embedding model Please write relations as disease, relation, and symptom in CSV format. If a disease or symptom consists of multiple instances, don't use a comma, write them as separate triples as shown in the format. You should also handle abbreviations. replace abbreviations with full terms. Don't use other words different from "associated with", "caused by", "causes" and "characterized by" as the relation."

Dil modellerinin verilen cümlelerden üçlüleri elde etme başarılarını değerlendirmek için 100 adet manuel olarak hazırlanan altın standart cümle üçlüsünden oluşan bir referans veri seti kullanılmıştır. Her bir LLM tarafından üretilen üçlülerin kalitesi, bu referans veri setiyle karşılaştırılarak değerlendirilmiştir. Bu yöntem, her bir LLM'nin üçlü üretme kabiliyetinin niceliksel olarak değerlendirilmesini sağlamıştır. Değerlendirmeler sonucunda GPT-4 modelinin daha başarılı olduğu görülmüştür.

GPT-4 yardımıyla elde edilen üçlüler bir ön işleme aşamasına tabi tutulmuştur. Tanımlanmamış ilişkilere sahip olan, üçten fazla varlık içeren veya eksik değerlere sahip üçlüler veri setinden kaldırılmıştır. Bunun ardından, Scispacy kütüphanesinden `en_ner_bc5cdr_md` NER modeli ve Semptom Ontolojisi kullanılarak hastalık, semptom ve kimyasal varlık üçlüleri grafik veri setine entegre edilmiştir. Kimyasal varlıkların da veri setine dahil edilmesi hastalıklar ve semptomlar arasındaki ilişkilerin daha iyi anlaşılmaısına ve yeni potansiyel bağlantıların ortaya çıkarılmasına katkı sağlamaktadır. Veri setinde bulunan varlık türleri "is a" ilişkisi ile üçlü veri setine dahil edilmiştir. Örneğin, "COVID-19,is a,disease", "fever, is a, symptom" ve "SARS-CoV-2 virus, is a, chemical" gibi üçlüler bu yapıya eklenmiştir. Bu yaklaşım, her bir varlığın semantik olarak kategorize edilmesini sağlar, bu da bilgi grafiğinin yapısını ve kapsamını iyileştirir.

Hastalık-Semptom Graf Veri Setine İlişkin Bilgiler: Elde edilen üçlülerden oluşturulan veri setinde ön işlem aşamalarından sonra 2649 adet benzersiz üçlü elde edilmiştir. Toplam beş adet ilişki türü bulunmaktadır. Bu ilişki

türleri ve sayıları sırasıyla "characterized by", 204 adet, "associated with", 1108 adet, "caused by", 252 adet, "causes" 17 adet ve "is a", 1045 adettir. Veri seti toplam 875 adet hastalık, 139 adet semptom ve 54 adet kimyasal varlık ismi içermektedir. Oluşturulan graf setinden örnek bir kesit Şekil 5.4'de, tüm veri setinin graf yapısındaki temsili Şekil 5.5'de gösterilmiştir. Şekil 5.6'da ise oluşturulan veri setinden örnek bir graf gösterimi sunulmuştur. Graf veri setine ait istatistiksel bilgiler Tablo 5.4 ve Tablo 5.5 sunulmuştur.

entity1_med	relation	entity2_med
chronic obstructive pulmonary disease	associated with	gastroesophageal reflux disease
gastroesophageal reflux disease	associated with	chronic cough
gastroesophageal reflux disease	associated with	laryngitis
gastroesophageal reflux disease	associated with	chest pain
wheezing	associated with	nocturnal cough
lower respiratory tract infection	associated with	laryngeal cleft
lower respiratory tract infection	associated with	laryngeal cleft
allergic rhinitis	associated with	sinusitis
allergic rhinitis	associated with	conjunctivitis
allergic rhinitis	associated with	eczema
allergic rhinitis	associated with	eustachian tube dysfunction
allergic rhinitis	associated with	otitis
reflux (lpr)	caused by	inflammatory reaction
status epilepticus	caused by	aminophylline
airway disorders	associated with	chronic cough
allergic rhinitis	associated with	anxiety
allergic rhinitis	associated with	depression
neurodevelopmental disorder	associated with	depression
neurodevelopmental disorder	associated with	anxiety
hyperthermia	associated with	muscle rigidity
hyperthermia	associated with	tachycardia

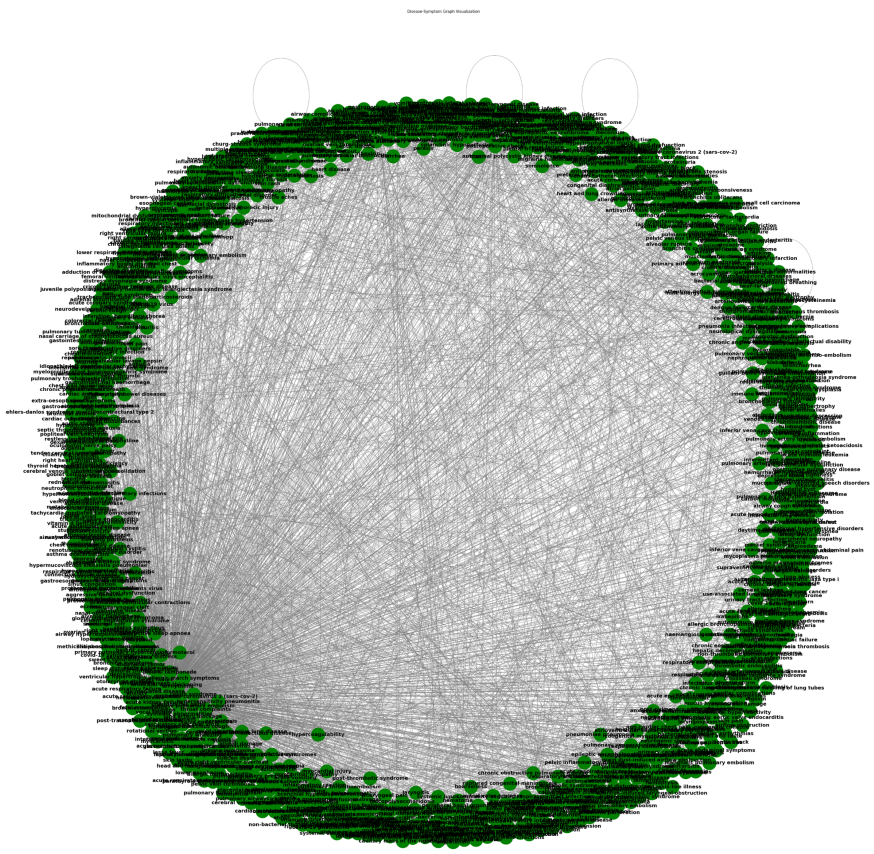
Şekil 5.4: Graf setinden bir örnek bir bölüm

Tablo 5.4: Verisetindeki İlişki Türleri

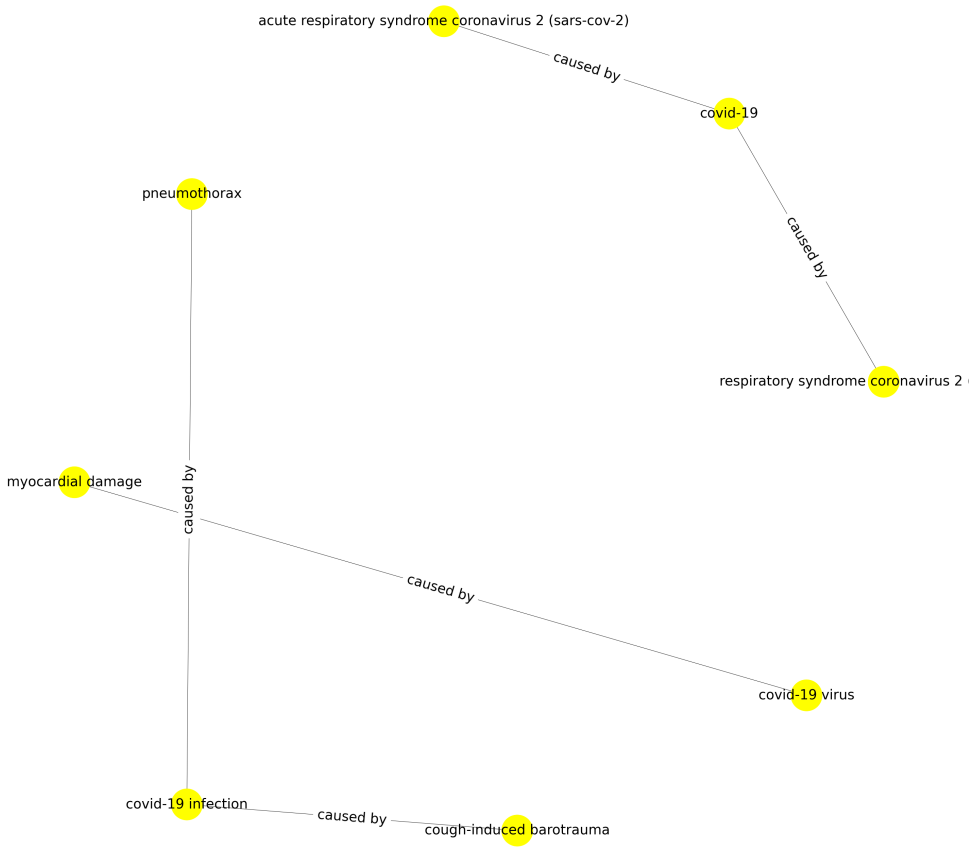
İlişki Türü	**Sayı**
Characterized by	204
Associated with	1108
Caused by	252
Causes	17
Is a	1045
Toplam	**2.626**

Tablo 5.5: Verisetindeki Varlık Türleri

Varlık Türü	**Sayı**
DISEASE	875
SYMPTOM	139
CHEMICAL	54
Toplam	**1.068**



Şekil 5.5: Veri setinin graf yapısında temsili



Şekil 5.6: Verisetinden örnek bir graf gösterimi

Graf Tamamlama Yöntemlerinin Uygulanması:

Bu çalışmada, inşa edilen bilgi grafı üzerinde potansiyel/bilinmeyen bağlantıları keşfetmek için TransE, ComplEx, DistMult ve HolE olmak üzere dört farklı graf tamamlama yöntemi kullanılmıştır. Veri seti yüzde 80 eğitim ve yüzde 20 test olarak iki bölüme ayrılmıştır. Model eğitimleri Python programlama dili kullanılarak Anaconda Navigator platformunda, GeForce RTX 3060 grafik kartı üzerinde yapılmıştır.

Graf tamamlama modellerinin performansını optimize etmek amacıyla, her bir yöntem için hiperparametre en iyileme süreci gerçekleştirilmiştir. Her bir model için en iyi olası hiperparametrelerin seçilmesini sağlamak amacıyla bir parametre uzayı oluşturulduktan sonra Amligraph 2.0 kütüphanesinin

select_ best _ model yöntemi ile en iyi parametre kombinasyonu seçilmiştir (Costabello et al., 2019). Bu yöntem ile rastgele aramalar algoritması kullanılarak 500 adet deneme sonucunda en iyi parametreler tespit edilmiştir. Bu parametrelere ilişki bilgilere Tablo 5.6'da yöntemlere göre en iyileme sonucu elde edilen değerler ise Tablo 5.7'de yer verilmiştir.

Tablo 5.6: Parametre Aralıkları ve Açıklamaları

Parametre Adı	**Açıklama**	**Değer Aralığı**
batches_count	Toplu iş sayısı	50, 100, 200, 300, 400
seed	Rastgele tohum değeri	0
epochs	Eğitim devirleri	20, 30, 40
k	Gömme boyutu	150, 300, 400, 500
eta	Öğrenme oranı	5, 10, 20, 50
loss	Kayıp fonksiyonu	nll, multiclass_nll
loss_params.margin	Marj değeri	0.5, 20
loss_params.alpha	Alpha değeri	0.5
embedding_model_params	Gömme model parametreleri	all
regularizer	Düzenleyici	LP
regularizer_params.p	Düzenleyici norm tipi	2
regularizer_params.lambda	Düzenleyici lambda değeri	0.00002, 0.0001, 0.00001
optimizer	Optimizasyon yöntemi	adam
optimizer_params.lr	Öğrenme oranı	[0.0001, 0.01]
verbose	Ayrıntılı çıktı modu	True

6 DENEYSEL SONUÇLAR

6.1 Semantik Benzerlik Sonuçları

Bu tezde, belirlenen solunum yolu hastalıkları için hastalık-semptom ilişkileri üç ayrı benzerlik yöntemi ile hesaplanarak, benzerlik skorlarına göre ayrı ayrı sıralanmıştır. Yöntemlerin başarıları her bir hastalık için bilinen semptomların üst sıralarda tespit edilebilme oranına göre değerlendirilmiştir. bu yöntemlerin hastalıklara göre doğrulanan semptom sayıları, 50, 100 ve 200 semptom sınırına göre incelenmiştir. Kosinüs, öklid ve nokta çarpım benzerliği yöntemlerinin ortalama değerlendirme puanları sırasıyla 0,61, 0,62 ve 0,66'dır.

Bu sonuçlara göre, Nokta Çarpımı benzerliği yöntemi en yüksek ortalama

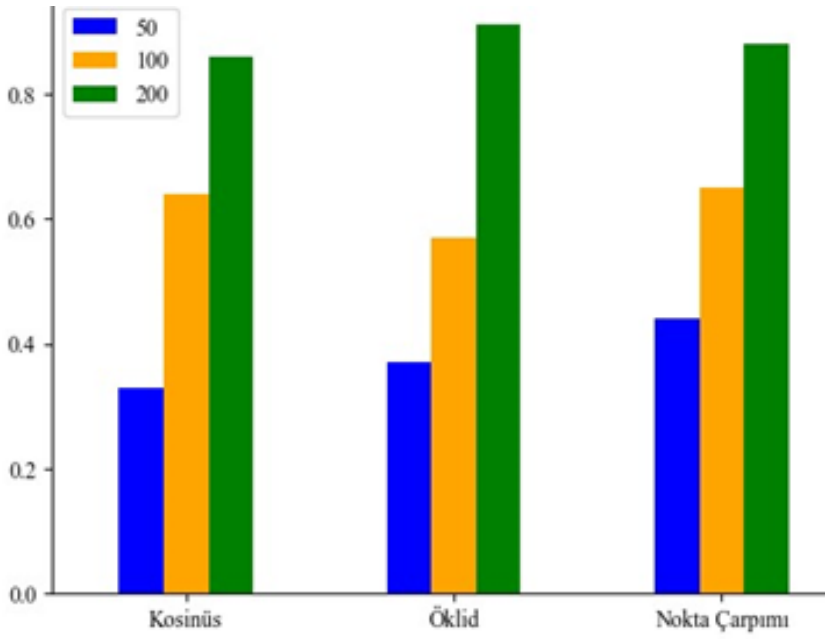
başarı skoruna (0,66) sahip olarak öne çıkmaktadır. Bu durum, Nokta Çarpımı yönteminin, hastalıklarla ilişkili semptomları sıralamada ve tespitte diğer yöntemlere göre daha iyi performans sergilediğini göstermektedir. Her bir hastalık için en alakalı semptomlar (Şekil 6.2, 6.4, 6.6, 6.8) ve nadir olarak nitelendirilebilecek sıralamaların sonunda yer alan 10 semptom (Şekil 6.3, 6.5, 6.7, 6.9) görselleştirilerek sunulmuştur. Bu semptomlar nokta çarpımı yöntemine göre yapılan sıralamalardan elde edilmiştir.

Tablo 5.7: En iyi Hiper Parametreler

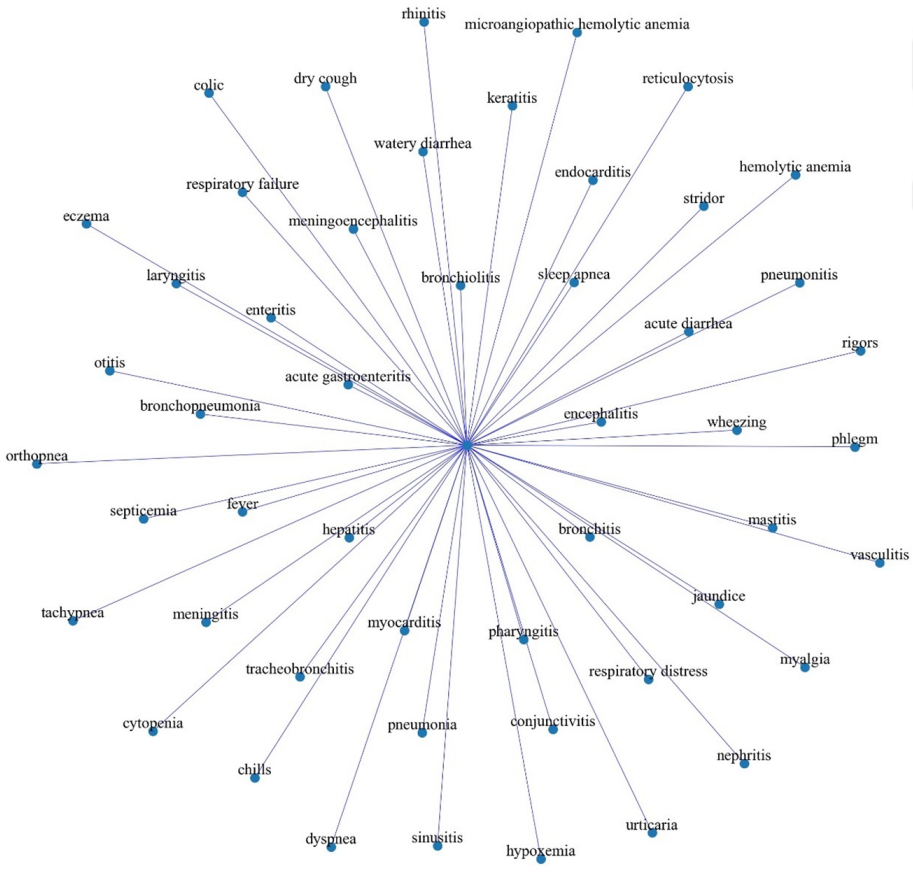
Model	**Parametreler**
TransE	batches count: 50, seed: 0, epochs: 20, k: 400, eta: 5, loss: nll, loss params: margin: 20.0, alpha: 0.5, negative corruption entities: all, regularizer: LP, regularizer params: p: 2, lambda: 2e-05, optimizer: adam, lr: 0.0001
ComplEx	batches count: 400, seed: 0, epochs: 30, k: 500, eta: 5, loss: multiclass nll, loss params: margin: 0.5, alpha: 0.5, negative corruption entities: all, regularizer: LP, regularizer params: p: 2, lambda: 0.0001, optimizer: adam, lr: 0.0001
DistMult	batches count: 300, seed: 0, epochs: 40, k: 400, eta: 5, loss: multiclass nll, loss params: margin: 0.5, alpha: 0.5, negative corruption entities: all, regularizer: LP, regularizer params: p: 2, lambda: 0.0001, optimizer: adam, lr: 0.0001
HolE	batches count: 300, seed: 0, epochs: 40, k: 400, eta: 10, loss: nll, loss params: margin: 20.0, alpha: 0.5, negative corruption entities: all, regularizer: LP, regularizer params: p: 2, lambda: 1e-05, optimizer: adam, lr: 0.0001

Tablo 6.1: Benzerlik Yöntemleri İçin Değerlendirme Sonuçları

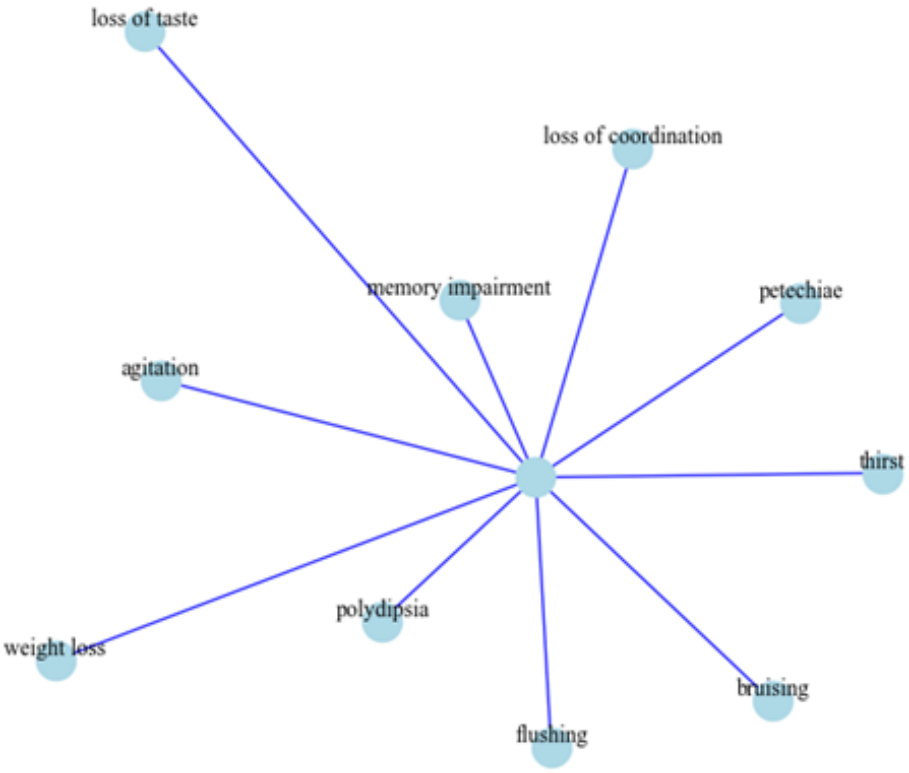
Benzerlik Yöntemi	**Hastalık**	**Doğrulanan Semptom Sayısı**	**50**	**100**	**200**
Cosine Similarity	COVID-19	18	4	9	17
	Bronchitis	10	5	9	10
	Asthma	12	5	8	9
	Pulmonary Embolism	12	2	6	9
Euclidean Similarity	COVID-19	18	6	11	16
	Bronchitis	10	4	7	10
	Asthma	12	5	6	11
	Pulmonary Embolism	12	4	6	10
Dot Product Similarity	COVID-19	18	6	9	17
	Bronchitis	10	7	10	10
	Asthma	12	6	7	10
	Pulmonary Embolism	12	3	6	9



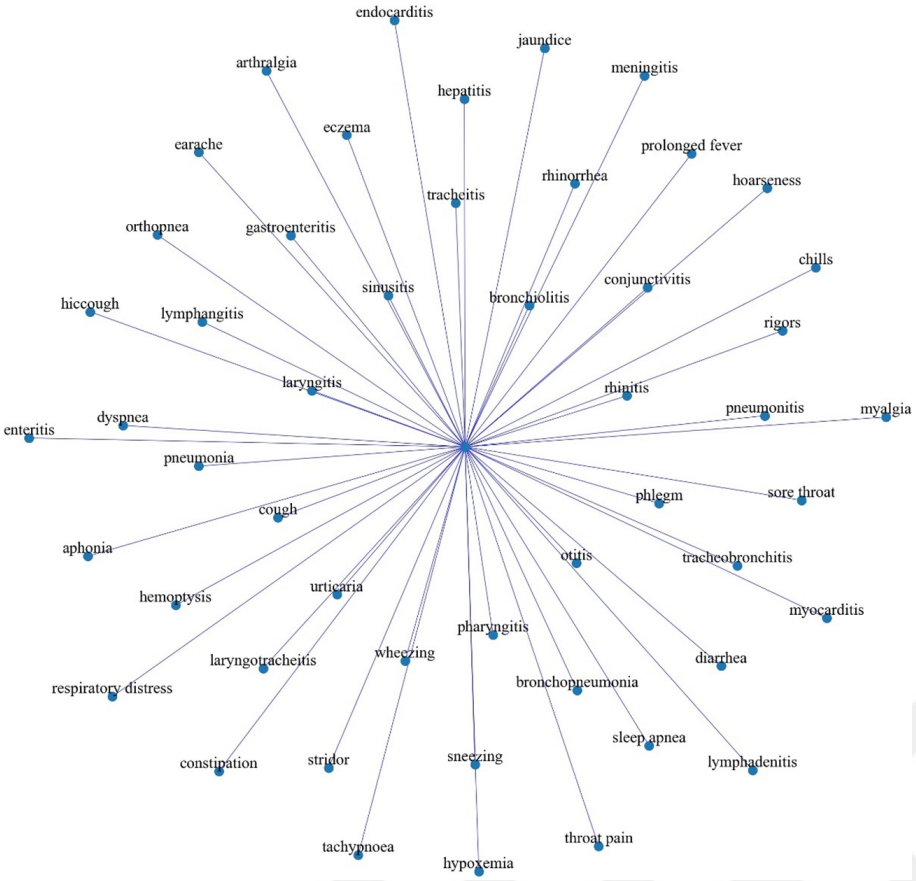
Şekil 6.1: Benzerlik Yöntemleri İçin Değerlendirme Sonuçlarının Ortalaması



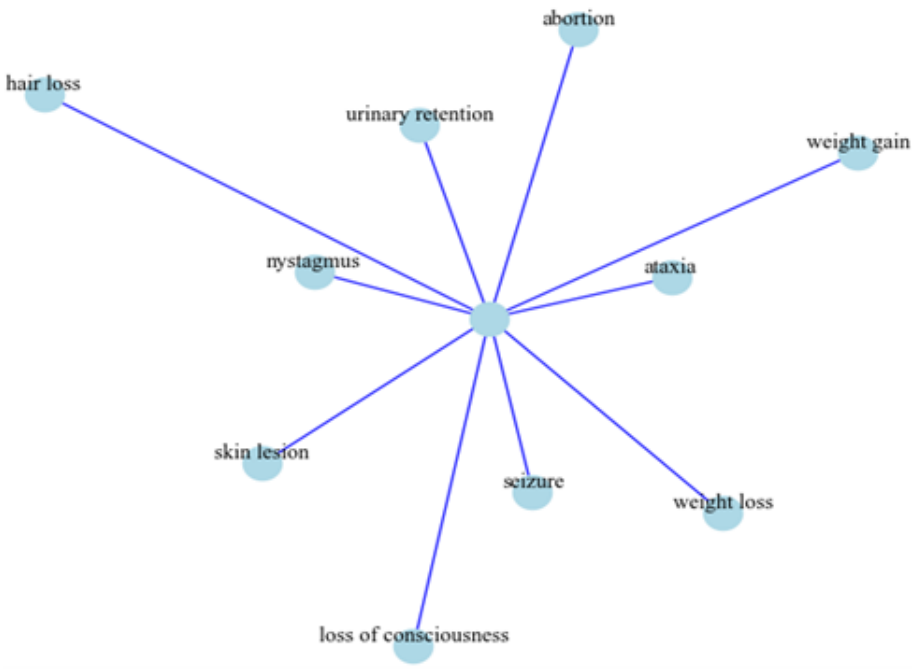
Şekil 6.2: COVID-19 ile İlişki Derecesi En Yüksek 50 Semptom



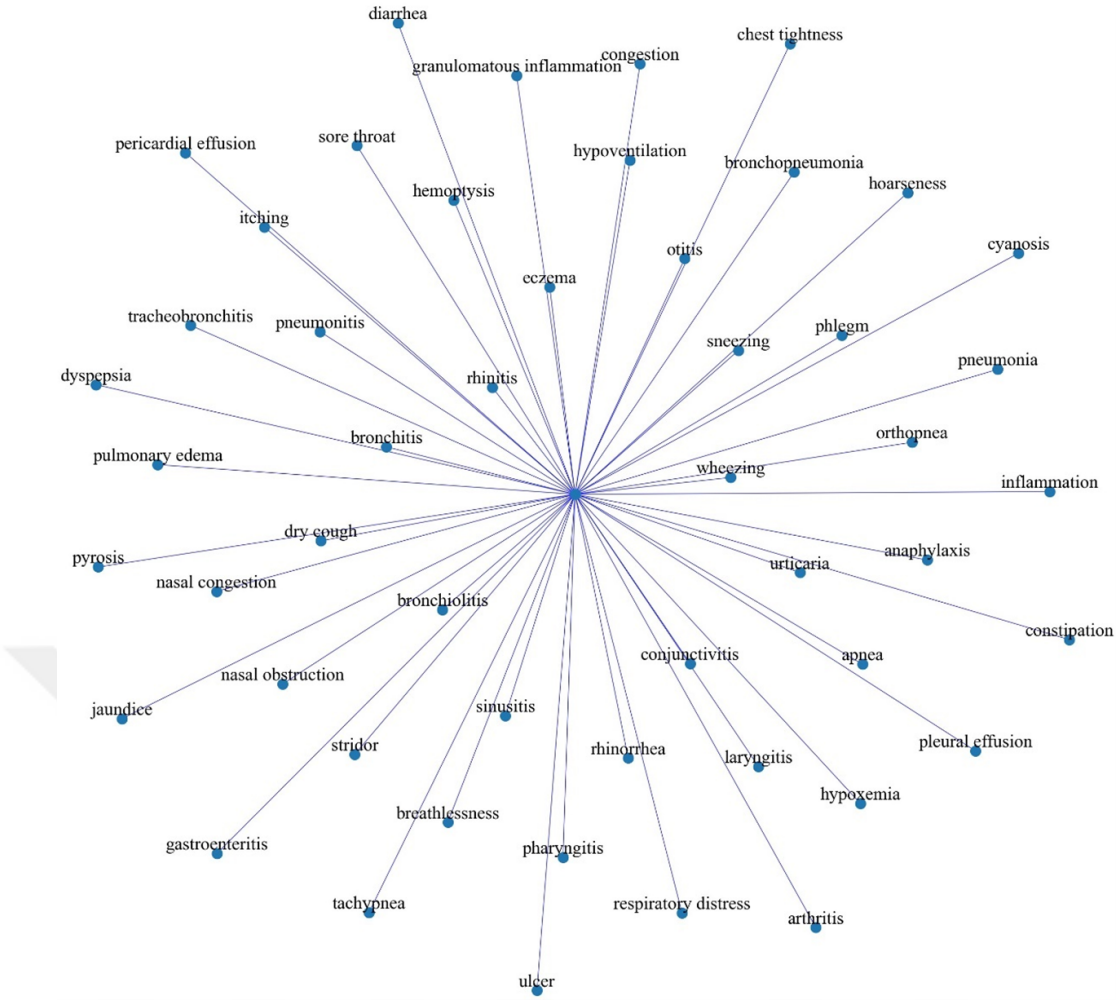
Şekil 6.3: COVID-19 ile İlişki Derecesi En Düşük 10 Semptom



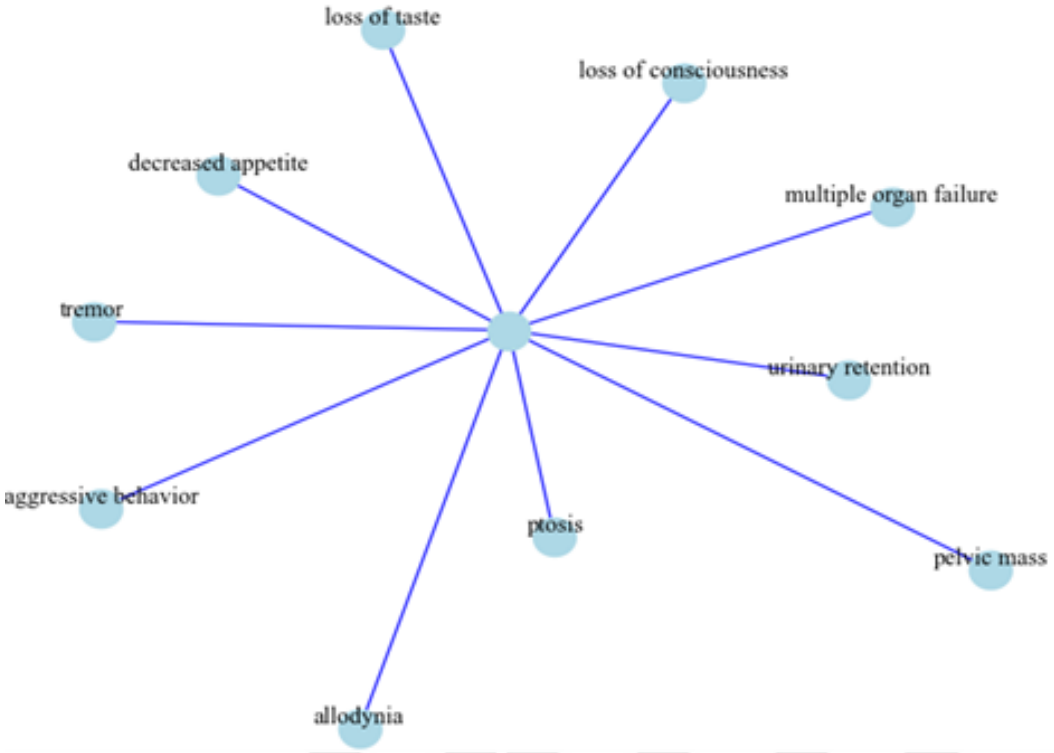
Şekil 6.4: Bronşit ile İlişki Derecesi En Yüksek 50 Semptom



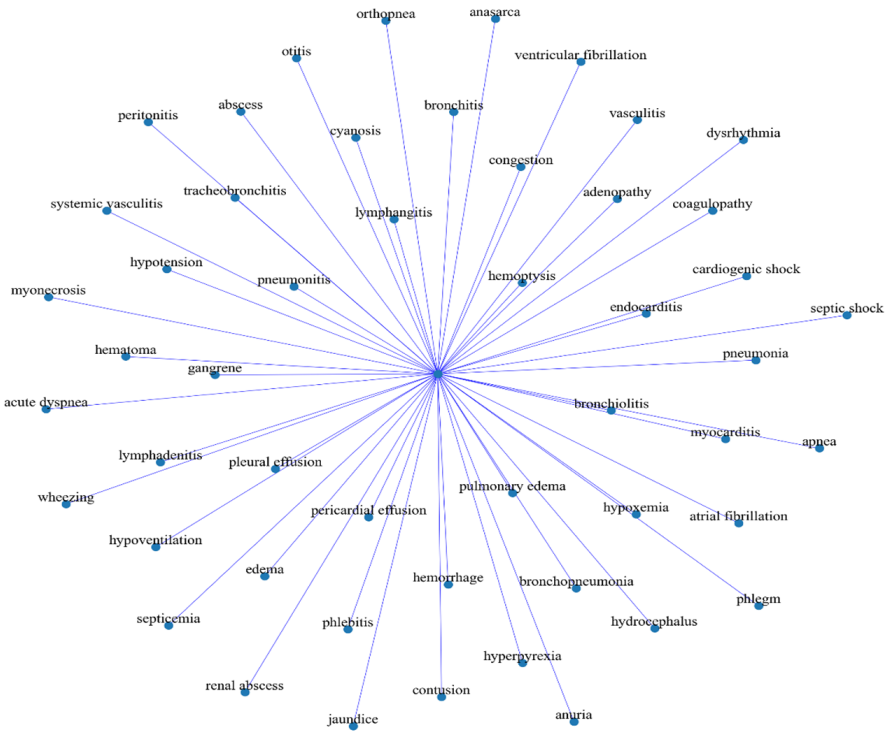
Şekil 6.5: Bronşit ile İlişki Derecesi En Düşük 10 Semptom



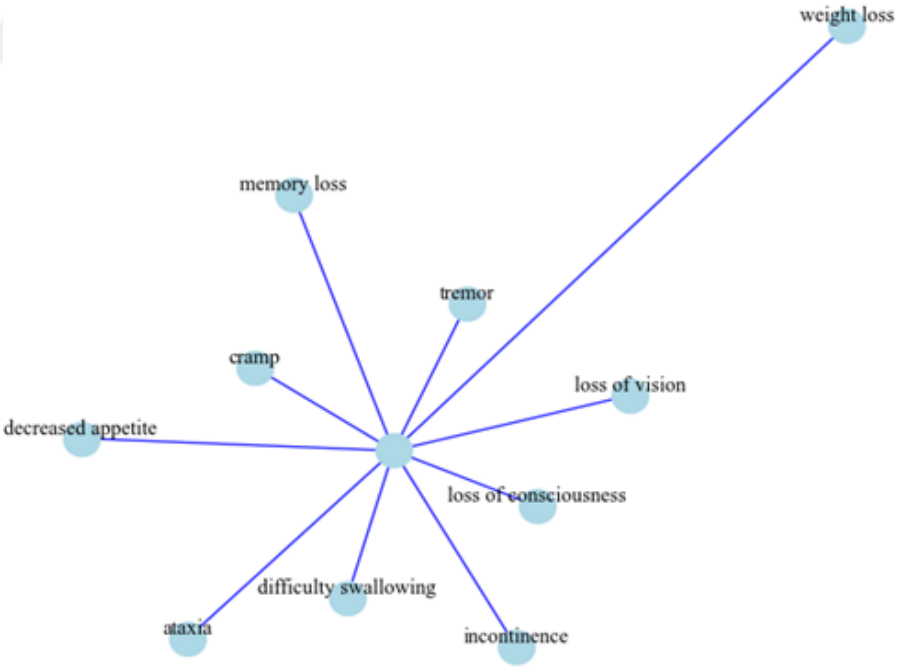
Şekil 6.6: Astım ile İlişki Derecesi En Yüksek 50 Semptom



Şekil 6.7: Astım ile İlişki Derecesi En Düşük 10 Semptom



Şekil 6.8: Pulmoner Emboli ile İlişki Derecesi En Yüksek 50 Semptom



Şekil 6.9: Pulmoner Emboli ile İlişki Derecesi En Düşük 10 Semptom

6.2 Graf Tamamlama Sonuçları

Bu bölümde, bilgi grafı oluşturma aşamasında ve graf tamamlama görevleri için gerçekleştirilen çalışmaların (ilişkilerin cümle düzeyinde sınıflandırılması, üçlülerin oluşturulması ve bağlantıların tahmin edilmesi) sonuçları sunulmaktadır.

6.2.1 Cümle Düzeyinde İlişki Sınıflandırma Sonuçları

Makalelerden elde edilen cümlelerin hastalık-semptom ilişkisi içerme durumlarına göre sınıflandırılmıştır. Tablo 6.2, cümle düzeyi ilişki sınıflandırmada kullanılan dört farklı modelinin performans metriklerini sunmaktadır: BERT-Base-multilingual-cased, BioBERT, SciBERT ve PubMedBERT. Elde edilen sonuçlar transformer tabanlı sınıflandırıcı modellerin ilişki sınıflandırmada yüksek düzeyde hassasiyet, hatırlama, F1 skoru ve doğruluk sergilediğini göstermiştir. Ancak, PubMedBERT tüm metriklerde en yüksek skorlara sahip olarak üstün performansını göstermektedir. Bu, PubMedBERT'in adından da anlaşılacağı gibi Pubmed makalelerin işlenmesinde uzmanlaşmış olduğunu ve bu nedenle daha iyi performans gösterdiğini göstermektedir. Modellerin performans metriklerindeki hafif farklılıklar, eğitim verileri, mimari veya modellerin ince ayar yapıldığı özel görevler nedeniyle oluşmuş olabilir.

Tablo 6.2: Performance metrics for various classification models

Classification Model	**Precision**	**Recall**	**F1-score**	**Accuracy**
BERT-Base-multilingual-cased	0.87	0.87	0.87	0.88
BioBERT	0.86	0.86	0.86	0.86
SciBERT	0.87	0.87	0.87	0.87
PubMedBERT	0.90	0.89	0.89	0.90

6.2.2 Üçlü Üretme Sonuçları

Hastalık-semptom ilişki içeren aday cümlelerden hastalık, ilişki, semptom üçlülerinin üretiminden büyük dil modellerinden faydalanılmıştır. Bu modellerin başarıları manuel olarak üçlü üretilen 100 cümle üzerinden accuracy (doğruluk), precision (kesinlik), recall (duyarlılık) ve F1-skor hesaplanarak

değerlendirilmiştir. Tablo 6.3, bilgi grafı oluşturmak için doğru üçlülerin üretimi konusunda Google LaMDA, GPT-3.5-turbo ve GPT-4 olmak üzere üç büyük dil modelinin performansını göstermektedir. Elde edilen sonuçlara göre GPT-4, tüm metriklerde diğer modelleri geride bırakarak, ilgili üçlüleri doğru bir şekilde üretme ve geri getirme konusunda daha iyi performans göstermiştir. GPT-3.5-turbo da özellikle duyarlılık ve F1 skoru açısından güçlü bir performans sergilemektedir, bu da modelin ilgili üçlüleri üretirken fazla sayıda alakasız üçlü üretmediğini düşündürebilir. Google LaMDA da oldukça iyi performans gösterse de, özellikle hatırlama ve F1 skoru açısından diğer modellerin gerisinde kalmaktadır.

Tablo 6.3: Performance metrics for Large Language Models

Büyük Dil Modeli	**Doğruluk**	**Kesinlik**	**Duyarlılık**	**F1-skor**
Google LaMDA	0.80	0.78	0.74	0.76
GPT-3.5-turbo	0.85	0.82	0.84	0.83
GPT-4	0.93	0.88	0.92	0.88

6.2.3 Graf Tamamlama Yöntemleri Sonuçları

Oluşturulan bilgi grafında potansiyel bağlantıları tahmin edebilmek için TransE, ComplEx DistMult ve HolE yöntemleri ile model eğitimleri gerçekleştirilmiştir. Elde edilen sonuçlar ve ilgili yöntemlerin literatürde kullanılan diğer veri setleri üzerinde karşılaştırma sonuçları Tablo 6.4'da verilmiştir. TransE modelinin daha başarılı olduğu görülmüştür. TransE'nin en düşük MR, en yüksek MRR ve isabet oranlarına sahip olması, hastalık-semptom bilgi grafındaki eksik bağlantıları veya düğümleri doğru tahmin etmede daha etkili kılmaktadır. Elde edilen sonuçlar bir veri kümesindeki üçlülerin sayısının, farklı graf tamamlama modellerinin performansını etkileyebileceğini ortaya koymaktadır. Daha büyük veri kümeleri, modeller arasındaki performans farklılıklarını azaltabilirken, daha küçük veri kümeleri, belli modellerin güçlü ya da zayıf yönlerini daha net bir şekilde ortaya çıkarabilir.

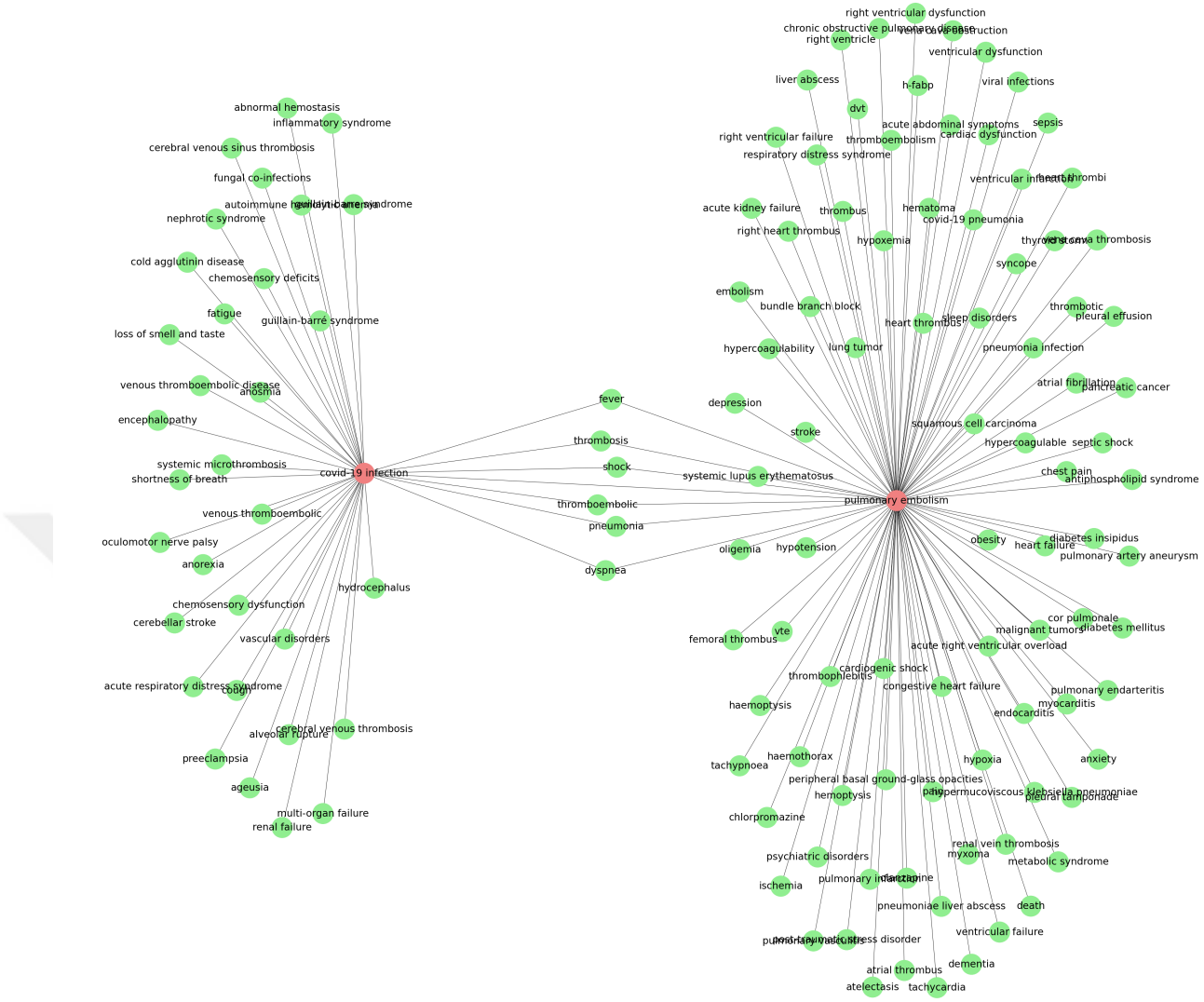
Tablo 6.4: Graf Tamamlama Yöntemlerinin Başarı Karşılaştırması

Veriseti	**Üçlü Sayısı**	**Yöntem**	**MR**	**MRR**	**Hit@1**	**Hit@3**	**Hit@10**
FB15K-237	310,116	TransE	211	0.31	0.22	0.34	0.48
		ComplEx	197	0.31	0.21	0.34	0.49
		DistMult	211	0.30	0.21	0.33	0.48
		HolE	190	0.30	0.21	0.33	0.48
WN18-RR	93,003	TransE	3143	0.22	0.03	0.38	0.52
		ComplEx	4229	0.50	0.47	0.52	0.58
		DistMult	4832	0.47	0.43	0.48	0.54
		HolE	7072	0.47	0.44	0.49	0.54
YAGO3-10	1,179,040	TransE	1210	0.50	0.41	0.56	0.67
		ComplEx	3153	0.49	0.40	0.54	0.65
		DistMult	2301	0.48	0.39	0.53	0.64
		HolE	7525	0.47	0.38	0.52	0.62
DS-KG	2,625	TransE	237	0.24	0.18	0.27	0.34
		ComplEx	453	0.09	0.06	0.10	0.16
		DistMult	486	0.08	0.05	0.09	0.14
		HolE	601	0.08	0.06	0.08	0.12

6.2.4 İlişki Analizi Görselleştirmeleri

Python'daki Networkx kütüphanesi kullanılarak hastalık-semptom bilgi grafı için görsel analizler gerçekleştirilmiştir. Bu sayede bilgi grafının farklı kullanım alanlarına dair öneriler sunulmuştur.

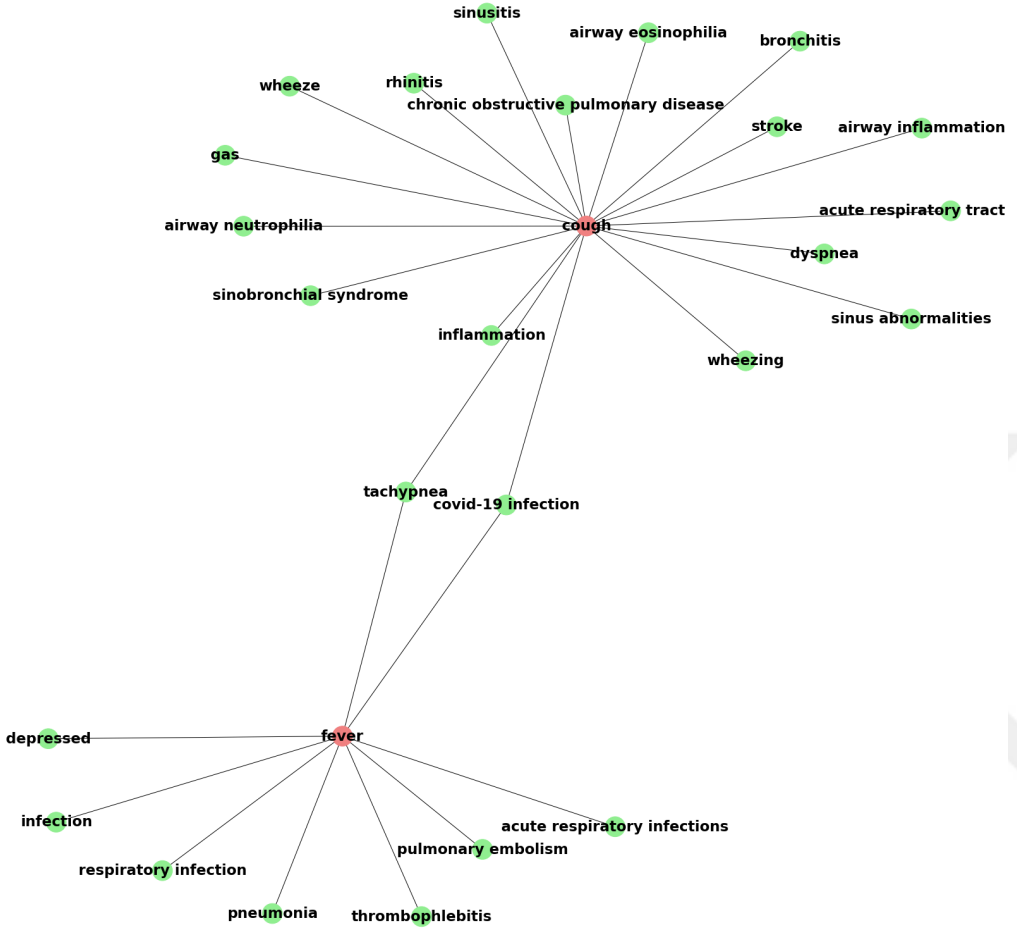
Oluşturulan bilgi grafı kullanılarak hastalıklar arasındaki ortak semptomlar analiz edilebilir. Özellikle solunum yolu hastalıkları gibi ortak semptomlar içeren hastalık gruplarında bu analizler hastalıkların karakteristik semptomlarını belirlemede oldukça etkilidir. Şekil 6.10, COVID-19 ve pulmoner emboli ile ilişkili semptomları ve hastalıkları analiz ederek bu hastalıklar arasındaki ortak ve ayırt edici bağlantıları ortaya çıkarmaktadır.



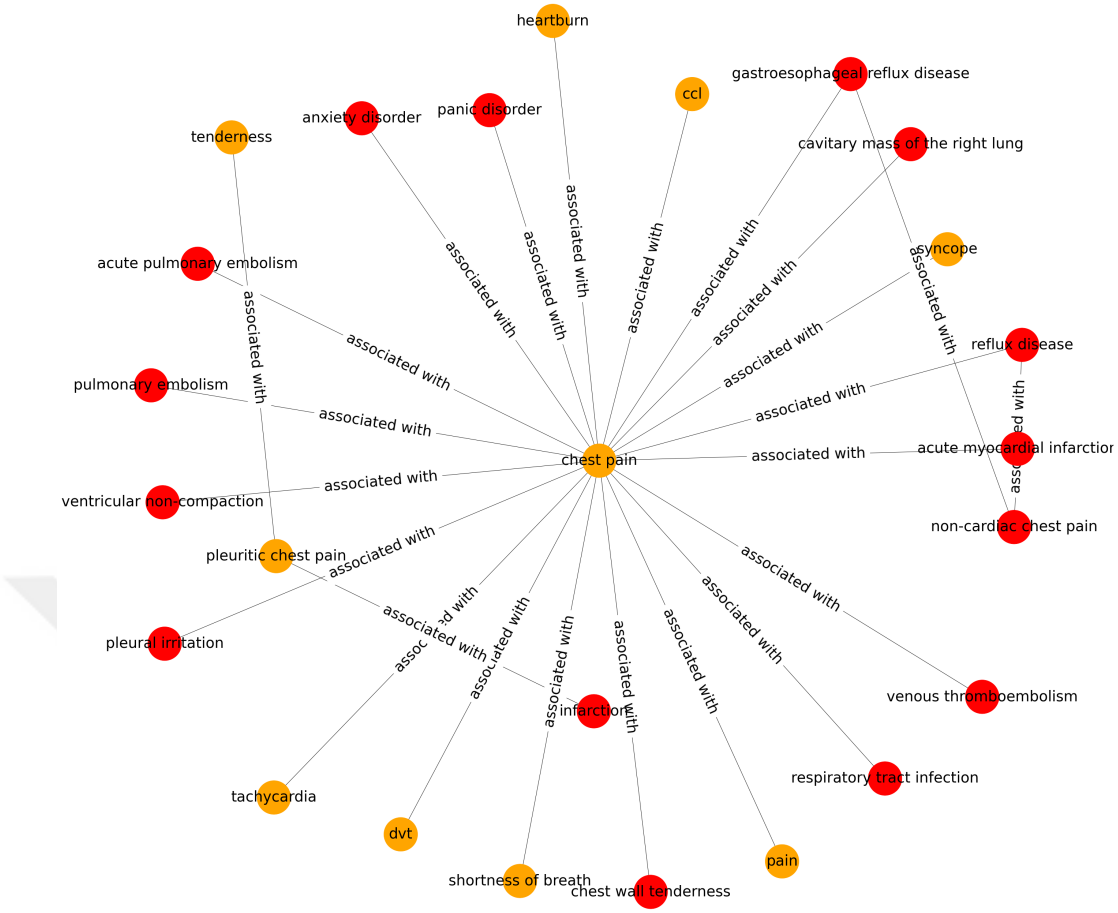
Şekil 6.10: COVID-19 ve Pulmoner Emboli Bağlantıları (Kırmızı Düğümler: Hedef Hastalık, Yeşil Düğümler: İlişkili Hastalık veya Semptomlar).

Ayrıca, bir semptomun ilişkili olabileceği diğer hastalıklara kolayca ulaşılabilir. Örneğin, göğüs ağrısı gibi bir çok hastalığa işaret edebilecek bir semptomun bağlantılı olabileceği diğer hastalıklar bilgi grafı sayesinde analiz edilebilir. Şekil 6.12, 'göğüs ağrısının' diğer hastalıklar ve semptomlarla olası ilişkilerini göstermektedir. Şekil 6.11'da, öksürük ve ateş semptomları

birlikte değerlendirilerek bunlarla ilişkili olabilecek diğer hastalıklar tespit edilmektedir.



Şekil 6.11: Ateş ve öksürük bağlantıları. (Kırmızı Düğüm: Hedef Belirtileri, Yeşil Düğümlerle İlişkili Hastalık veya Belirtiler)



Şekil 6.12: Göğüs Ağrısıyla İlişkili Belirtiler ve Hastalıklar (Kırmızı Düğümler: Hastalık, Sarı Düğümler: Belirtiler)

6.2.5 Yeni Bağlantı Tahmini Sonuçları

TransE graf tamamlama modeli kullanılarak oluşturulan bilgi grafındaki potansiyel hastalık-semptom ilişkisi içeren yeni bağlantılar keşfedilmiştir. Bağlantı tespit etme yöntemleri doğru bağlantıları ilk 100 sırada tespit edebilme oranlarına göre karşılaştırılmıştır. Önerilen yöntem kapsamında varlık sıklığı (*entity frequency* ve küme üçgenleri (*cluster triangles*) yöntemi hibrit olarak kullanıldığında doğru bağlantıların tespit edilme oranının arttığı görülmektedir. Sonuçlar Tablo 6.5'de yer almaktadır.

Tablo 6.5: İlişki Türleri ve Yöntemlerin Başarım Sonuçları

İlişki Türü	**Yöntem**	**Toplam Tahmin Sayısı**	**Doğrulanan Tahmin Sayısı**	**Accuracy**
Characterized By	Varlık Sıklığı	15	7	0.47
	Graf Derecesi	28	12	0.42
	Küme Üçgenleri	33	21	0.63
	Önerilen Yöntem	10	8	0.80
Associated With	Varlık Sıklığı	16	9	0.56
	Graf Derecesi	35	25	0.71
	Küme Üçgenleri	34	32	0.94
	Önerilen Yöntem	11	11	1.00

Tahmin edilen bağlantılardan pulmoner emboli hastalığı ile ilişkili olarak venous thromboembolism ve respiratory distress tespit edilmiştir. Tespit edilen yeni bağlantılar graf yapısına eklenerek bilgi grafının kapsamı genişletilebilir. Şekil 6.13'te yeni eklenen bu bağlantılar görülebilir. Bu bağlantıların bilimsel literatürden tespit edilmesi ve yapısal formatta sunulması hastalıkların daha kapsamlı bir biçimde incelenmesine ve anlaşılmasına imkan tanıyarak tıp ve sağlık bilimleri alanında yenilikçi araştırmaların önünü açar. Ayrıca, yeni bağlantılar ve örüntüler, bilim insanlarına ve sağlık profesyonellerine hastalıkların nedenlerini ve gelişim yollarını daha iyi anlama fırsatı sunarak tedavi stratejilerinin geliştirilmesine katkı sağlar.

Şekil 6.13: Pulmoner Emboli bağlantıları (Kırmızı Düğüm: Hedef Hastalık, Yeşil Düğümler: İlişkili Hastalık veya Belirtiler, Mavi Düğümler: Yeni Tahmin Edilen İlişkiler).

7 SONUÇ VE TARTIŞMA

Bu tezde, medikal bilgi çıkarımı kapsamında bilimsel literatürden hastalıklarla ilgili ilişkileri tespit ederek bunları yapısal bir formatta sunmayı amaçlayan çeşitli yöntemler geliştirilmiştir. Seçilen hastalıklar arasında COVID-19, bronşit, astım ve pulmoner emboli gibi solunum yolu hastalıkları bulunmaktadır. Tez kapsamında, PubMed'den elde edilen özgün bir veri seti kullanılmıştır. İlişki çıkarımı için semantik benzerlik ve graf embedding yöntemleri kullanılarak çeşitli analizler yapılmış ve yeni yaklaşımlar ortaya

konmuştur.

Semantik benzerlik yöntemleri kapsamında, hastalık ve semptomlar arasındaki vektörel mesafe hesaplaması kosinüs, öklid ve nokta çarpımı yöntemleri kullanılarak gerçekleştirilmiştir. Deneysel sonuçları, nokta çarpım benzerliği kullanılarak hastalıklar ve ilgili semptomları tespit etmede diğer yöntemlere kıyasla daha iyi performans sergilediğini göstermiştir. Ek olarak, düşük benzerlik değerlerine sahip olmasına rağmen hastalıklarla ilişkili olan nadir semptomlar da analiz edilmiştir.

Bu tezde, graf yöntemlerini kullanarak ilişki analizleri yapabilmek için bilimsel literatürden türetilen özel bir bilgi grafı oluşturulmuştur. Bildiğimiz kadarıyla, solunum yolu hastalıklarına özgü, literatür tabanlı spesifik bir bilgi grafı daha önce geliştirilmemiştir. Bilgi grafının oluşturulma aşamasında karşılaşılan alt problemler için çeşitli çözüm önerileri geliştirilmiştir. Makale özetlerinden uygun cümlelerin seçimi için kural ve transformer tabanlı olmak üzere iki aşamalı bir sınıflandırma mimarisi oluşturulmuştur. Cümlelerden ilişki üçlülerinin üretiminde büyük dil modelleri kullanılmıştır. Elde edilen graf üzerinde Graf tamamlama yöntemleri ile inşa edilen bilgi grafı üzerinde embedding tabanlı TransE, DistMult, ComplEx ve HolE modelleri eğitilerek optimizasyonlar yapılmıştır. Son olarak potansiyel bağlantıları en doğru şekilde tespit edebilmek için varlık sıklığı ve küme üçlüsüne dayanan hibrit bir yöntem önerilerek tespit edilen yeni bağlantıların güvenilirliği arttırılmıştır. Elde edilen bağlantılar GPT-4 modeli ile otomatik olarak doğrulanmıştır.

Elde edilen sonuçlar, PubMedBERT ve GPT-4 gibi modellerin ilişki sınıflandırması ve üçlüler üretiminde önemli avantajlar sunduğunu ortaya koymaktadır. Bu modellerin biyomedikal literatürdeki varlıkları ve ilişkileri doğru bir şekilde tanımlama ve sınıflandırma yeteneği, kapsamlı bir bilgi grafının geliştirilmesi için oldukça önemlidir. Ayrıca, graf embedding tekniklerinin ve TransE gibi graf tamamlama modellerinin uygulanması, potansiyel ilişkileri ortaya çıkarma ve bilgi grafiğindeki eksik bilgileri doldurma konusunda umut verici bir yol sunmaktadır.

Bu yaklaşımlar, tıp profesyonelleri ve araştırmacılar için zaman ve kaynak

tasarrufu sağlamanın yanı sıra, hastalıklar ve semptomlar arasındaki karmaşık ilişkileri yapılandırılmış bir formata dönüştürmektedir. Tıbbi araştırma alanı gelişmeye devam ettikçe, biyomedikal bilgiler ile ileri hesaplama tekniklerinin entegrasyonu giderek daha önemli hale gelecektir. Bu çalışmanın yöntemleri ve sonuçları tıbbi araştırma ve bilgi yayılımının geleceğinde doğal dil işleme ve metin madenciliğine dayalı derin öğrenme yöntemlerinin önemini vurgulamaktadır.

Kaynaklar

PUBMED. Available at: https://pubmed.gov.tr. Accessed: December 10, 2023.

PubMed Overview. Available at: https://pubmed.ncbi.nlm.nih.gov/about/#:~:text=PubMed%20Overview,health%E2%80%93both%20globally%20and%20personally. Accessed: December 10, 2023.

MEDLINE PubMed Production Statistics. Available at: https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html. Accessed: December 10, 2023.

Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2020). Attention guided capsule networks for chemical-protein interaction extraction. Journal of Biomedical Informatics, 103, 103392.

Peng, Y., Rios, A., Kavuluru, R., & Lu, Z. (2018). Extracting chemical–protein relations with ensembles of SVM and deep learning models. Database.

Zhou, H., Liu, Z., Ning, S., Lang, C., Lin, Y., & Du, L. (2019). Knowledge-aware attention network for protein-protein interaction extraction. Journal of Biomedical Informatics, 96, 103234.

Zhou, H., Li, X., Yao, W., Liu, Z., Ning, S., Lang, C., & Du, L. (2019). Improving neural protein-protein interaction extraction with knowledge selection. Computational Biology and Chemistry, 83, 107146.

Onye, S. C., Akkeleş, A., & Dimililer, N. (2018). RelSCAN–a system for extracting chemical-induced disease relation from biomedical literature. Journal of Biomedical Informatics, 87, 79-87.

Deng, Y., Xu, X., Qiu, Y., Xia, J., Zhang, W., & Liu, S. (2020). A multimodal deep learning framework for predicting drug–drug interaction events. Bioinformatics, 36 (15), 4316-4322.

Feng, Y. H., Zhang, S. W., & Shi, J. Y. (2020). DPDDI: a deep predictor for drug-drug interactions. BMC Bioinformatics, 21 (1), 1-15.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. In Proceedings of the AMIA Symposium (p. 17). American Medical Informatics Association.

Leaman, R., Wei, C. H., & Lu, Z. (2013). DNorm: Disease Name Normalization with Pairwise Learning to Rank. Bioinformatics.

Wei, C. H., Kao, H. Y., & Lu, Z. (2013). PubTator: A Web-based Text Mining Tool for Assisting Biocuration. Nucleic Acids Research.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association: JAMIA.

Fukuda, K. I., Tsunoda, T., Tamura, A., & Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. In Pacific Symposium on Biocomputing (pp. 707-718).

Çelikten, A., Onan, A., & Bulut, H. (2022, May). Investigation of Biomedical Named Entity Recognition Methods. In The International Conference on Artificial Intelligence and Applied Mathematics in Engineering (pp. 218-229). Cham: Springer International Publishing.

Tamames, J. (2005). Text detective: a rule-based system for gene annotation in biomedical texts. BMC Bioinformatics, 6, 1-8.

Kazama, J., Makino, T., & Ohta, Y. (2007). Tuning support vector machines for biomedical named entity recognition. In Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain (Vol. 2002, No. 1-8, p. 05).

Zhou, G. (2004). Recognizing names in biomedical texts using hidden markov model and SVM plus sigmoid. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) (pp. 1-7).

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 34(8), 1381-1388.

Wu, G., Tang, G., Wang, Z., Zhang, Z., & Wang, Z. (2019). An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition. IEEE Access, 7, 113942-113949.

Wei, H., Gao, M., Zhou, A., Chen, F., Qu, W., Wang, C., & Lu, M. (2019). Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF. IEEE Access, 7, 73627-73636.

McDonald, R., & Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics, 6, 1-7.

Batbaatar, E., & Ryu, K. H. (2019). Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach. International Journal of Environmental Research and Public Health, 16(19), 3628.

Wei, H., Gao, M., Zhou, A., Chen, F., Qu, W., Wang, C., & Lu, M. (2019). Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF. IEEE Access, 7, 73627-73636.

Scepanovic, S., Martin-Lopez, E., Quercia, D., & Baykaner, K. (2020). Extracting medical entities from social media. In Proceedings of the ACM Conference on Health, Inference, and Learning (pp. 170-181).

Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific Data, 6(1), 52.

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

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2020). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Proceedings of the 18th BioNLP Workshop and Shared Task.

Wu, C. S., Kuo, C. J., Su, C. H., Wang, S. H., & Dai, H. J. (2020). Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. Journal of Affective Disorders, 260, 617-623.

Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing and Applications, 34(1), 721-744.

Eisman, A. S., Shah, N. R., Eickhoff, C., Zerveas, G., Chen, E. S., Wu, W. C., & Sarkar, I. N. (2020). Extracting angina symptoms from clinical notes using pre-trained transformer architectures. In AMIA Annual Symposium Proceedings, 2020, 412. American Medical Informatics Association.

Leiter, R. E., Santus, E., Jin, Z., Lee, K. C., Yusufov, M., Chien, I., ... & Lindvall, C. (2020). Deep natural language processing to identify symptom documentation in clinical notes for patients with heart failure undergoing cardiac resynchronization therapy. Journal of Pain and Symptom Management, 60(5), 948-958.

Wang, J., Abu-el-Rub, N., Gray, J., Pham, H. A., Zhou, Y., Manion, F. J., ... & Zhang, Y. (2021). COVID-19 SignSym: a fast adaptation of a general clinical NLP tool to identify and normalize COVID-19 signs and symptoms to OMOP common data model. Journal of the American Medical Informatics Association, 28(6), 1275-1283.

Lybarger, K., Ostendorf, M., Thompson, M., & Yetisgen, M. (2021). Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. Journal of Biomedical Informatics, 117, 103761.

Zhou, X., Menche, J., Barabási, A.-L., et al. (2014). Human symptoms–disease network. Nature Communications, 5, 4212.

Hassan, M., Makkaoui, O., Coulet, A., & Toussaint, Y. (2015). Extracting disease-symptom relationships by learning syntactic patterns from dependency graphs. In BioNLP 15, 184.

Abulaish, M., & Parwez, M. A. (2019). DiseaSE: A biomedical text analytics system for disease symptom extraction and characterization. Journal of Biomedical Informatics, 100, 103324.

Zlabinger, M., Hofstätter, S., Rekabsaz, N., & Hanbury, A. (2020). DSR: A Collection for the Evaluation of Graded Disease-Symptom Relations. In European Conference on Information Retrieval, Springer, Cham, 433-440.

Wada, S., Iida, R., Torisawa, K., Takeda, T., Manabe, S., & Matsumura, Y. (2018). Symptom Extraction and Disease-Symptom Relation Recognition from Web Texts with Multi-Column Convolutional Neural Networks.

Pechsiri, C., & Piriyakul, R. (2022). Construction of Disease-Symptom Knowledge Graph from Web-Board Documents. Applied Sciences, 12(13), 6615.

Geleta, D., Nikolov, A., Edwards, G., Gogleva, A., Jackson, R., Jansson, E., et al. (2021). Biological Insights Knowledge Graph: an integrated knowledge graph to support drug development. Biorxiv, 2021-10.

Huang, Z., Hu, Q., Liao, M., Miao, C., Wang, C., & Liu, G. (2021). Knowledge Graphs of Kawasaki Disease. Health Information Science and Systems, 9, 1–8.

Gao, M., Lu, J., & Chen, F. (2022). Medical Knowledge Graph Completion Based on Word Embeddings. Information, 13(4), 205.

Ebeid, I. A., Hassan, M., Wanyan, T., Roper, J., Seal, A., & Ding, Y. (2021). Biomedical Knowledge Graph Refinement and Completion Using Graph Representation Learning and Top-K Similarity Measure. In Proceedings of the 16th International Conference on Diversity, Divergence, Dialogue (iConference 2021), Beijing, China, March 17–31, 2021, Part I, Vol. 16. Springer International Publishing, 112–123.

Rossanez, A., Dos Reis, J. C., Torres, R. D. S., & de Ribaupierre, H. (2020). KGen: A Knowledge Graph Generator from Biomedical Scientific Literature. BMC Medical Informatics and Decision Making, 20(4), 1–24.

Socrates, V. (2022). Extraction of Diagnostic Reasoning Relations for Clinical Knowledge Graphs. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 413–421.

Dai, Z., Wang, X., Ni, P., Li, Y., Li, G., & Bai, X. (2019, October). Named entity recognition using bert bilstm crf for chinese electronic health records. In 2019 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI) (pp. 1-5). IEEE.

Sousa, D., Lamurias, A., & Couto, F. M. (2020). Using neural networks for relation extraction from biomedical literature. In Artificial Neural Networks (pp. 289-305). New York, NY: Springer US.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, s.282–289.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ...

& Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Lee, M. (2023). A mathematical investigation of hallucination and creativity in gpt models. Mathematics, 11(10), 2320.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240), 1-113.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). Palm 2 technical report. arXiv preprint arXiv:2305.10403.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems, Vol. 26.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In International Conference on Machine Learning. PMLR, 2071-2080.

Yang, B., Yih, W. T., He, X., Gao, J., & Deng, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575.

Nickel, M., Rosasco, L., & Poggio, T. (2016). Holographic embeddings of knowledge graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30.

Ma, X., Conrad, T., Alchikh, M., Reiche, J., Schweiger, B., & Rath, B. (2018). Can we distinguish respiratory viral infections based on clinical features? A prospective pediatric cohort compared to systematic literature review. Reviews in Medical Virology, 28(5), e1997.

Van der Sar, I. G., Wijsenbeek, M. S., Braunstahl, G. J., Loekabino, J. O., Dingemans, A. C., In 't Veen, J. C. C. M., & Moor, C. C. (2023).

Differentiating interstitial lung diseases from other respiratory diseases using electronic nose technology. Respiratory Research, 24(1), 271.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Disease Ontology Consortium. [n.d.]. Human Disease Ontology. Available at: http://disease-ontology.org. Accessed: January 7, 2024.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Kentaro Inui, Jing Jiang, Vincent Ng, & Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3615-3620.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., others, & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

Costabello, L., Pai, S., Le Van, C., McGrath, R., McCarthy, N., & Tabacof, P. (2019). AmpliGraph: a Library for Representation Learning on Knowledge Graphs. DOI: 10.5281/zenodo.2595043. Available at: https://doi.org/10.5281/zenodo.2595043.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ...

& Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Lee, M. (2023). A mathematical investigation of hallucination and creativity in gpt models. Mathematics, 11(10), 2320.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240), 1-113.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). Palm 2 technical report. arXiv preprint arXiv:2305.10403.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems, Vol. 26.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In International Conference on Machine Learning. PMLR, 2071-2080.

Yang, B., Yih, W. T., He, X., Gao, J., & Deng, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575.

Nickel, M., Rosasco, L., & Poggio, T. (2016). Holographic embeddings of knowledge graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30.

Ma, X., Conrad, T., Alchikh, M., Reiche, J., Schweiger, B., & Rath, B. (2018). Can we distinguish respiratory viral infections based on clinical features? A prospective pediatric cohort compared to systematic literature review. Reviews in Medical Virology, 28(5), e1997.

Van der Sar, I. G., Wijsenbeek, M. S., Braunstahl, G. J., Loekabino, J. O., Dingemans, A. C., In 't Veen, J. C. C. M., & Moor, C. C. (2023).

Differentiating interstitial lung diseases from other respiratory diseases using electronic nose technology. Respiratory Research, 24(1), 271.

Costabello, L., Pai, S., Le Van, C., McGrath, R., McCarthy, N., & Tabacof, P. (2019). AmpliGraph: a Library for Representation Learning on Knowledge Graphs. *March*, 2019. DOI: 10.5281/zenodo.2595043. Available at: `https://doi.org/10.5281/zenodo.2595043`

TEŞEKKÜR

Bu akademik yolculuk boyunca, her bir adımda tüm içtenlikleriyle bana rehberlik eden ve değerli bilgi ve deneyimleriyle çalışmama yön veren kıymetli danışmanlarım Prof. Dr. Hasan Bulut ve Prof. Dr. Aytuğ Onan'a; tez sürecimde ilham kaynağım olan, desteğini her zaman yanımda hissettiğim AKGÜN Teknoloji ailesi ve değerli yöneticim Kadir Budak'a; hayatımın her anında benimle olan sevgili aileme ve bana inanıp destek olan herkese en derin teşekkürlerimi sunarım.

12/07/2024

Azer Çelikten

ÖZGEÇMİŞ

Kişisel Bilgiler

Ad Soyad: Azer Çelikten

Eğitim Bilgileri

Ege Üniversitesi, Bilgisayar Mühendisliği (Doktora)

Gazi Üniversitesi, Bilgisayar Mühendisliği (Yüksek Lisans)

Muğla Sıtkı Koçman Üniversitesi, Bilgisayar Mühendisliği (Lisans)

İş Deneyimi

Akgün Teknoloji, Yapay Zeka Mühendisi (Nisan 2023, Devam Ediyor)

Manisa Celal Bayar Üniversitesi Yazılım Müh. Araştırma Görevlisi (2018-2023)

Ziraat Teknoloji, Yazılım Mühendisi (2015-2018)

Tez Döneminde Yapılan Akademik Yayınlar

Çelikten, A., Onan, A., & Bulut, H. (2024). *Utilizing Large Language Models for Constructing and Enriching Disease-Symptom Knowledge Graphs.* **Transactions on Intelligent Systems and Technology** (Under Review)

Çelikten, A., Onan, A., & Bulut, H. (2023). *A Semantic Similarity-Based Approach to Extract Respiratory Disease-Symptom Relations from Biomedical Literature.* **Journal of the Faculty of Engineering and Architecture of Gazi University** (Accepted)

Çelikten, A., Onan, A., & Bulut, H. (2022, May). *Investigation of Biomedical Named Entity Recognition Methods.* **The International Conference on Artificial Intelligence and Applied Mathematics in Engineering**, 218-229. Cham: Springer International Publishing.

Çelikten, A., & Bulut, H. (2021). *Turkish medical text classification using bert.* **2021 29th Signal Processing and Communications Applications Conference (SIU)**, 1-4.

Çelikten, A., Uğur, A., & Bulut, H. (2021). *Keyword extraction from biomedical documents using deep contextualized embeddings.* **2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)**, 1-5.