Journal of the Faculty of Engineering and Architecture of Gazi University 40:1 (2025) 121-134

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

A semantic Similarity-Based approach to extract respiratory disease-symptom relations from biomedical literature

Azer Çelikten[1,3]*, Hasan Bulut[1], Aytuğ Onan[2]
[1]Department of Computer Engineering, Faculty of Engineering, Ege University, 35100, İzmir, Türkiye
[2]Department of Computer Engineering, Faculty of Engineering and Architecture, İzmir Katip Çelebi University, 35620, İzmir, Türkiye
[3]Akgün Technology, 06930, Ankara, Türkiye

Highlights:

- Symptom recognition method using medical word vectors and symptom ontology
- Sorting symptoms by semantic similarity
- Identifying rare symptoms with low scores

Keywords:

- Biomedical Named Entity Recognition
- Disease-Symptom Relation Extraction
- Biomedical Text Mining
- Respiratory Diseases and Symptoms

Article Info:
Research Article
Received: 13.09.2023
Accepted: 06.01.2024

DOI:
10.17341/gazimmfd.1354324

Correspondence:
Author: Azer Çelikten
e-mail: azer.celikten@gmail.com
phone: +90 534 014 8884

Graphical/Tabular Abstract

This study proposes a method for extracting disease-symptom relationships, aimed at identifying rare symptoms not mentioned in healthcare resources that could be associated with diseases, and evaluating the extent of relation between diseases and symptoms. The proposed method for disease-symptom relation extraction is showed in Figure A.

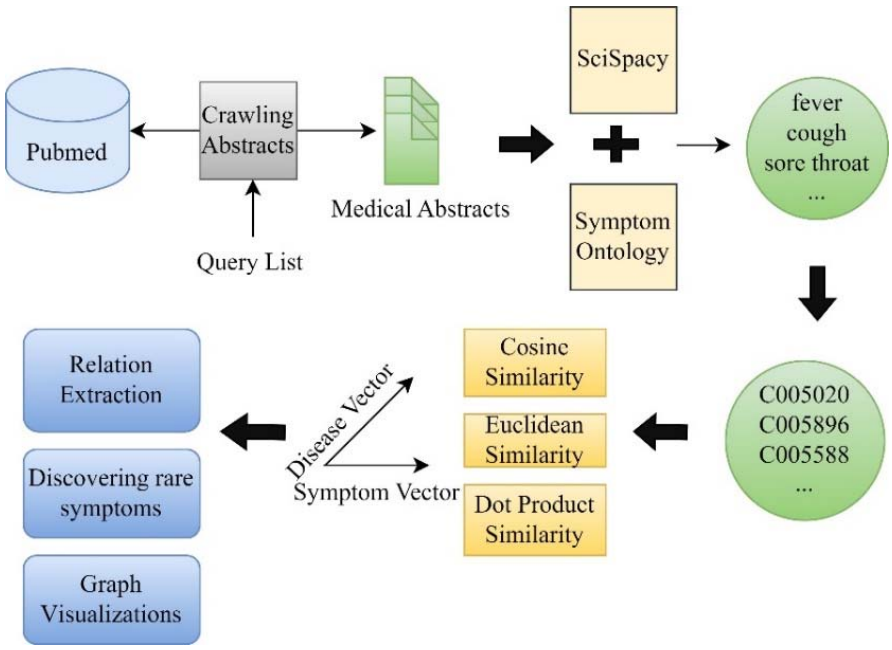


Figure A. Proposed Disease-Symptom Relation Extraction Method

Purpose: The purpose of this research is to develop a disease-symptom relation extraction method for the early diagnosis of respiratory diseases.

Theory and Methods: We introduce a hybrid named entity recognition approach to extract disease and symptom names from medical texts and performed different similarity-based approaches to capture relations between these entities.

Results: Our method was evaluated on a dataset of respiratory diseases, including asthma, bronchitis, pulmonary embolism, and coronavirus diseases. We discovered the rare symptoms that are low relation score but related to the disease and demonstrated the efficacy of similarity-based approaches in establishing connections between diseases and symptoms. Our results show that the dot product similarity approach is outperformed for capturing disease and symptom relationships.

Conclusion: In conclusion, our research introduces an innovative disease-symptom relation extraction method that leverages advanced natural language processing techniques. This method has the potential to improve early diagnosis and patient care in the field of respiratory diseases.

Journal of the Faculty of Engineering and Architecture of Gazi University 40:1 (2025) 121-134

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Biyomedikal literatürden solunum yolu hastalıkları ve semptom ilişkilerinin çıkarılması için semantik benzerlik temelli bir yaklaşım

Azer Çelikten[1,3]*, Hasan Bulut[1], Aytuğ Onan[2]
[1]Ege Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 35100, Bornova, İzmir, Türkiye
[2]İzmir Katip Çelebi Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, 35620, İzmir, Türkiye
[3]Akgün Teknoloji, 06930, Ankara, Türkiye

ÖNEÇIKANLAR

- Tıbbi kelime vektörleri ve semptom ontolojisini kullanarak semptom tanıma yöntemi
- Semptomları semantik benzerlikle sıralama
- Düşük skorlu nadir semptomları belirleme

Makale Bilgileri
Araştırma Makalesi
Geliş: 13.09.2023
Kabul: 06.01.2024

DOI:
10.17341/gazimmfd.1354324

Anahtar Kelimeler:
Bilgi çıkarımı, biyomedikal varlık ismi tanıma, biyomedikal ilişki çıkarımı, hastalık-semptom ilişkileri, metin madenciliği

ÖZ
Biyomedikal alandaki artan makale sayısıyla birlikte, hastalıklar ve semptomlar hakkında keşfedilen değerli bilgiler akademik literatürde saklı kalmaktadır. Biyomedikal metinleri işlemek ve doğal dil işleme ve metin madenciliği yöntemlerini kullanarak bu bilgileri çıkarmak, erken teşhis, klinik karar destek sistemleri geliştirmek ve ontolojileri güncellemek için oldukça önemlidir. Solunum yolu hastalıklarının ateş, öksürük, nefes darlığı gibi birçok ortak semptomları olduğundan, hastalıkları semptomlara göre ayırt etmek, hastalığın erken evrelerinde doğru teşhisi sağlar. Bu çalışmada, sağlık kaynaklarında belirtilmeyen ancak hastalıkla ilişkili olabilecek nadir semptomları belirlemek ve hastalıkların semptomlarla ilişki derecesini tespit etmek için bir hastalık-semptom ilişkisi çıkarma yöntemi önerilmiştir. İlk olarak, tıbbi metinlerdeki hastalıkları ve semptomları tanımlamak için hibrit bir varlık ismi tanıma yöntemi önerilmiştir. Sonrasında, semptomlar ve hastalıklar normalize edilerek, semptomların hastalıklarla ilişki dereceleri semantik benzerlik skorlarına göre sıralanmıştır. Önerilen yöntem solunum yolu hastalıklarından oluşan özgün bir veri seti üzerinde değerlendirilmiştir. Bu veri seti, astım, bronşit, pulmoner emboli ve koronavirüs hastalıklarına ait akademik makale özetlerinden oluşmaktadır. Sonuç olarak, karakteristik semptomlara ek olarak, sağlık kaynaklarında bahsedilmeyen ancak hastalıkla ilişkilendirilebilecek nadir semptomlar keşfedilmiştir. Önerilen yöntem ile hastalıkların semptomları arasındaki ilişkilerin tespitinde 0,66 ortalama benzerlik skoru ile nokta çarpımı benzerlik yönteminin daha başarılı olduğu görülmüştür. Nadir semptomların ise literatür değerlendirmesi yapılarak hastalıklar ile ilişkisi ortaya çıkarılmıştır.

A semantic Similarity-Based approach to extract respiratory disease-symptom relations from biomedical literature

HIGHLIGHTS

- Symptom recognition method using medical word vectors and symptom ontology
- Sorting symptoms by semantic similarity
- Identifying rare symptoms with low scores

Article Info
Research Article
Received: 13.09.2023
Accepted: 06.01.2024

DOI:
10.17341/gazimmfd.1354324

Keywords:
Information extraction, biomedical named entity recognition, biomedical relation extraction, disease-symptom relations, text mining

ABSTRACT
In the biomedical domain, the surge in article volume means valuable insights on diseases and symptoms are often hidden in academic literature. Leveraging natural language processing and text mining to sift through biomedical texts is vital for advancing early diagnosis, enhancing clinical decision support systems, and refining ontologies. Particularly for respiratory diseases, which share symptoms like fever, cough, and breathlessness, differentiating between diseases based on symptoms is crucial for early and accurate diagnosis. This study introduces a method for extracting disease-symptom relationships, aiming to identify rare symptoms not mentioned in health resources but potentially related to diseases, and to ascertain the association strength between diseases and symptoms. Initially, a hybrid entity recognition approach was proposed for identifying diseases and symptoms in medical texts. Then, the diseases and symptoms were normalized, and their associations ranked by semantic similarity scores. Evaluated on a dataset of respiratory diseases, including academic article abstracts on asthma, bronchitis, pulmonary embolism, and COVID-19, the study uncovered rare symptoms in addition to characteristic ones. The dot product similarity method proved more effective, achieving an average similarity score of 0.66, in establishing the associations between diseases and symptoms, revealing the significance of literature validation in identifying rare symptom-disease relations.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *azer.celikten@gmail.com , hasan.bulut@ege.edu.tr , aytug.onan@ikcu.edu.tr / Tel: +90 534 014 8884

1. Giriş (Introduction)

Solunum yolu veya akciğer hastalıkları, Avrupa Birliği istatistiklerine göre önemli bir ölüm oranına neden olmaktadır [1]. Bu hastalıkların yüksek prevalansı ve morbiditesi, yüksek hastaneye yatış oranları nedeniyle ilaç maliyetleri ve giderleri açısından sağlık sistemi için önemli maliyetlere yol açmaktadır. Bu nedenle, erken teşhis ve doğru tedavi, ölüm oranlarını ve sağlık maliyetlerini azaltmak için kritik öneme sahiptir. 2019 yılında ortaya çıkan Covid-19, yaygın solunum yolu hastalıklarının yanı sıra pulmoner komplikasyonlara da neden olabilmektedir [2]. Bu hastalıklarda ateş, öksürük ve nefes darlığı gibi benzer semptomlar ortak olarak gözlense de farklı patofizyolojik süreçlere sahiptirler ve hastalığın erken evrelerinde doğru teşhis ve uygun tedavi için ayırt edilmelidirler. Solunum hastalıkları ve semptomlar arasındaki ilişkilerin belirlenmesi, erken teşhis, klinik karar destek sistemlerinin geliştirilmesi, biyomedikal bilgi grafikleri oluşturulması ve ontolojilerinin zenginleştirilmesi için kritik öneme sahiptir.

Tıptaki gelişmelerle birlikte solunum yolu hastalıkları ve semptomları hakkında yeni bilgilere ulaşılmaktadır. Hekimler ve tıp uzmanları, bu gelişmeleri PubMed [3] gibi biyomedikal veri tabanları aracılığıyla takip edebilirler. Ulusal Tıp Kütüphanesi tarafından yönetilen ve sağlık alanındaki bilimsel literatürü içeren bir kaynak olan PubMed 36 milyon alıntı ve makale özeti içermektedir [4]. PubMed'teki bilgiler yapılandırılmamış bir formatta sunulduğundan tıbbi gelişmelere ulaşmak, incelemek ve analiz etmek manuel arama yöntemleri ile gerçekleştirilmektedir. PubMed istatistiklerine göre bilimsel literatür araştırmaları kapsamında 2022 yılında, toplamda 2,58 milyar arama gerçekleştirilmiştir [5]. Bu nedenle, büyük hacimli ve önemli bilgiler içeren PubMed makaleleri içerisinde yer alan değerli bilgilerin yapılandırılması, analiz edilmesi ve birbirleriyle ilişkilendirilmesi ihtiyacı doğmuştur. Bilginin manuel olarak çıkarılması, içerdiği büyük hacim nedeniyle zaman alıcı, pahalı ve verimsiz bir çözümdür. Doğal dil işlemedeki gelişmeler, metinlerden bilgi çıkarmak ve aralarındaki ilişkileri bulmak için verimli çözümler sunar. Bilgi çıkarma iki adımdan oluşur. İlk adım, hastalıklar, semptomlar, ilaçlar, kimyasallar ve proteinler gibi medikal varlık adlarının belirlenme sürecidir. İkinci adım, çıkarılan varlıklar arasındaki belirli ilişkileri bulmaktır. Metinlerden kişi, kuruluş ve yer isimleri gibi bazı isimleri çıkarmak için çeşitli yöntemler önerilmiş olsa da, biyomedikal alana özgü kelimelerin, karmaşık ifadelerin ve terimlerin aşırı kullanımı nedeniyle tıbbi metinlerden varlık ismi çıkarımı daha zordur. Biyomedikal varlık ismi çıkarımından sonra, bu varlıklar arasındaki ilişkilerin belirlenmesi uzman bilgisi ve etiketli veri gerektirdiğinden, ilişki çıkarımı süreci daha karmaşıktır.

Biyomedikal ilişki çıkarım alanındaki çalışmaların çoğu, araştırmaya açık etiketli veri kümelerinde faydalanılarak ilaç-protein [6, 7], protein-protein [8, 9], hastalık-kimyasal [10] ve ilaç-ilaç [11-13] etkileşimlerini tespit etmek için gerçekleştirilmiştir. Hastalık-semptom ilişkilerini belirlemeye yönelik etiketli veri kümesine ulaşmadaki kısıtlar ve hastalık-semptom adlarının birbiri yerine kullanılabildiğinden ayırt edilebilmesindeki zorluklar nedeniyle az sayıda çalışma mevcuttur. Biyomedikal alanda doğal dil anlama açısından hastalıklar ve semptomları birbirinden ayırmak ve aralarındaki ilişkiyi belirlemek önemli bir araştırma konusudur.

Bu çalışmada, tıbbi metinlerden semptomları çıkarmak ve hastalıklar ile semptomlar arasındaki ilişkileri belirlemek için bir yöntem önerilmiştir. Önerilen yöntem altı adımda gerçekleştirilmiştir. İlk adımda, PubMed'den belirli sorgular kullanılarak taranan astım, bronşit, pulmoner emboli ve Covid-19 hastalıklarına ait bilimsel makale özetlerini içeren bir solunum yolu hastalıkları veri seti oluşturulmuştur. İkinci adımda, metinlerdeki semptom ve hastalıkları tanımlamak ve ayırt edebilmek için hibrit bir biyomedikal varlık ismi tanıma yöntemi önerilmiştir. Üçüncü adımda, çıkarılan semptom ve hastalıklar UMLS kavram tanımlayıcıları ile eşleştirilmiştir. Son adımda, hastalıklar ve semptomların benzerlik skorlarına göre sıralı bir listesi oluşturularak hastalıklarla ilişkili olma potansiyeli olan nadir semptomlar belirlenmiştir.

Bu çalışmanın katkıları,

- Ön eğitimli medikal kelime vektörleri ve tıbbi ontolojiyi birleştiren hibrit semptom ismi tanıma yöntemi,
- Solunum yolu hastalık-semptom isimlerinin geçtiği biyomedikal makale özetlerini içeren bir özgün bir veri seti oluşturulması,
- Farklı semantik benzerlik yöntemleri ile hastalık-semptom ilişkileri analiz edilerek ilişki derecelerine göre semptomların sıralanması,
- İlişki skoru düşük olan ancak hastalıkla ilişkili olma potansiyeli olan nadir semptomların keşfedilmesidir.

2. Literatür Taraması (Literature Review)

Literatürde kullanılan semptom tanımlama yöntemleri sözlük tabanlı, kural tabanlı, makine öğrenimi tabanlı ve derin öğrenme tabanlı yöntemler olmak üzere 4 kategoride incelenebilir [14].

Sözlüğe dayalı yöntemler [15, 16], varlık türleri için kapsamlı ad listelerinden oluşan önceden tanımlanmış sözlükler gerektirir. Ek varlıkların olasılığı nedeniyle, sözlük tabanlı algoritmalar daha iyi doğrulukla ancak yeni terimlerin daha düşük tanınma oranlarıyla karakterize edilir. Sözlükler, genler ve hastalıklar gibi ortak biyolojik varlıkların adları için var olsalar bile, diğer biyomedikal terimler için yetersiz veya eksiktir.

Kural tabanlı yaklaşımı kullanan sistemlerde [17, 18], varlıklar, metin modellerine dayalı olarak manuel olarak tanımlanan kurallarla tanımlanır. Kural tabanlı yaklaşımlar, olası varlıkları keşfetmek için yapılandırılmış kural kalıplarını kullanır. Varlık türleri için kurallar oluşturmak, uzman bilgisi gerektiren bir süreçtir. Sözlüğe ve kurala dayalı yöntemlere dayalı çoğu geleneksel yaklaşım, uzman bilgisi içerdiği için güvenilirdir ancak büyük ölçüde iyi tanımlanmış sözlüklere ve manuel oluşturulmuş kurallara dayanmaktadır.

Makine öğrenimi tabanlı yöntemlerde, öznitelikler ve etiketlerden oluşan eğitim verileri kullanılarak modeller eğitilir ve test verileri kullanılarak modellerin performansı ölçülür [19]. Makine öğrenimi yaklaşımlarında, varlık ismi tanıma performansını etkileyen ana faktörler, manuel olarak oluşturulan öznitelikler ve kullanılan algoritmalardır. Öznitelik mühendisliği ve uzman bilgisi, başarılı sonuçlar elde etmek için önemlidir. Makine öğrenimine dayalı yaklaşımlar, probleme özgü veri kümelerinde iyi sonuçlar verir. Biyomedikal varlık ismi tanıma için, Gizli Markov Modelleri, Destek Vektör Makineleri ve Koşullu Rastgele Alanlar [20, 21] gibi çeşitli makine öğrenimi modelleri geliştirilmiştir.

Derin Öğrenme yöntemleri, varlık ismi tanıma ve diğer birçok doğal dil işleme görevini modellemek için çeşitli sinir ağı mimarilerini kullanır. Son zamanlarda, Uzun-Kısa Süreli Bellek ve Konvolüsyonel Sinir Ağları varlık ismi tanıma problemi için yaygın olarak kullanılmaktadır [22, 23]. Derin Öğrenme yöntemleri ve mevcut donanım sistemleri sayesinde çok büyük miktarda veri işlenebildiği gibi özellikler de otomatik olarak çıkarılmaktadır. Girdi olarak biyomedikal alan için özel olarak eğitilmiş kelime gömme modelleri kullanılarak, kelimeler vektör uzayında temsil edilir.

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. Jackson vd. doğal dil işleme yöntemlerini kullanarak taburcu raporlarında akıl hastalığı semptomlarını belirlemek için bir çalışma yürütmüştür [24]. Wu vd. ayrıca kural tabanlı ve makine öğrenimi modellerini kullanarak semptomları tanımlamak için zihinsel bozukluklarla ilgili bir elektronik sağlık kayıtları veri seti kullanmıştır [25]. Uddin vd. tarafından yapılan çalışmada, Uzun-Kısa Süreli Bellek yöntemi çevrimiçi halka açık bilgi kanalındaki metinlere uygulanarak depresif semptomlar incelenmiştir [26]. Eisman vd. [27] ve Leiter vd. [28], 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 vd. [29], 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 vd. [30], 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.

Metinlerden tıbbi varlıklar çıkarıldıktan sonra ikinci adım, tanımlanan varlık adları arasındaki ilişkilerin bulunmasıdır. Çoğu zaman, ilişkiler denetimli öğrenme, birlikte görülmeye dayalı istatistiksel yöntemler veya varlıklar arasındaki semantik benzerlik yoluyla bulunur [32]. 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. Hastalıkların ortaya çıkışına dayalı istatistiksel 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.

Zhou vd. tarafından gerçekleştirilen çalışmada PubMed bibliyografik kayıtlarını hastalık/semptom ile değerlendirerek kosinüs benzerliğini kullanarak hastalık-semptom ilişkileri çıkarılmıştır [33]. Hassan vd.., 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 [34]. Abulaish vd. iklime duyarlı hastalıklar için hastalık-hastalık, hastalık-semptom ve semptom-semptom ilişkilerini bulmak için bağımlılık ve sözdizimsel kalıpları kullanmışlardır [35]. Zlabinger vd. Tarafından sıralama yöntemlerinin performansını değerlendirmek için bir hastalık semptom koleksiyonu oluşturulmuştur. Hastalıklar ve semptomlar arasındaki ilişki, birlikte görülme istatistikleri kullanılarak belirlenerek, koleksiyon ile değerlendirilmiştir [36]. Wada vd. Q/A modülünden 120.000 cümleden oluşan veri setinde konvolüsyonel sinir ağı mimarisini kullanarak hastalıklar ve semptomlar arasındaki ilişkileri çıkarmışlardır [37].

Covid-19 gibi spesifik solunum yolu hastalıklarının semptomları üzerine yapılmış çalışmalara rağmen, metin madenciliği yöntemleri kullanılarak solunum yolu hastalıklarının semptomlarını grup olarak inceleyen bir çalışma bulunmamaktadır. Bu boşluk, bu alanda daha fazla araştırma için bir fırsat sunmaktadır. İlişkileri anlamak için benzerliğe dayalı yöntemler kullanan araştırmalar, tipik olarak kosinüs benzerliğini kullanmış, diğer semantik benzerlik yöntemleri ile ilgili bir çalışma gerçekleştirilmemiştir. Bunun yanı sıra, birçok ilişki çıkarma yöntemi, hastalıklar ile semptomlar arasındaki ilişkileri önceden belirlenmiş bilgilere dayandırır. Ancak, biyomedikal literatür, yaygın olarak bilinen semptomların yanı sıra, klinik vakalarda ortaya çıkan nadir semptomları da içerebilir. Bu çalışmada farklı benzerlik yöntemlerinin hastalık-semptom ilişkisi çıkarımındaki başarıları değerlendirilerek elde edilen semptomların hastalıklarla olan ilişki skorları belirlenmiştir. Ayrıca, semptom-hastalık biyomedikal bilgi grafları oluşturulmuş, semptomlar hastalıklara olan alaka derecelerine göre sıralanarak analiz edilmiş ve görselleştirilmiştir.

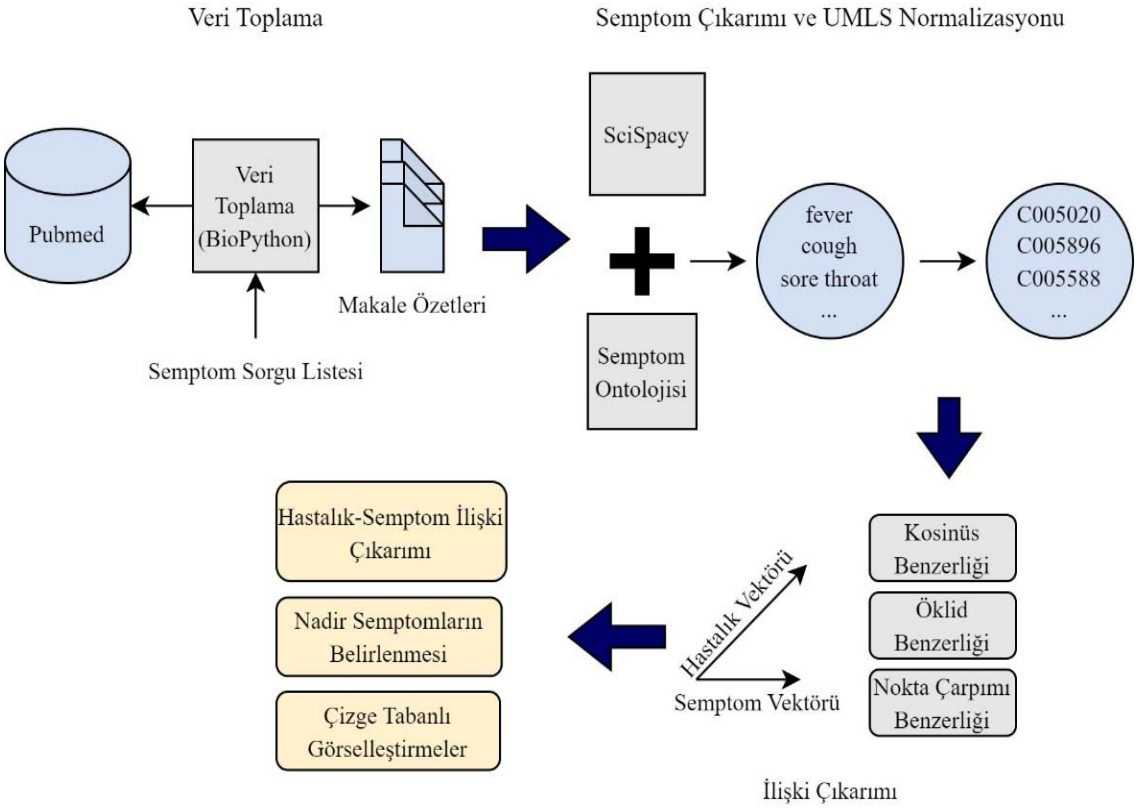
3. Yöntemler (Methods)

Biyomedikal makalelerden solunum yolu hastalıkları ve semptomları arasındaki ilişkileri belirlemek için önerilen yöntem akışı Şekil 1'de gösterilmiştir. Önerilen yöntem veri toplama, semptom çıkarımı, UMLS normalizasyonu, ilişki çıkarımı olmak üzere dört bölümden oluşmaktadır.

3.1. Veri seti (Dataset)

Bu çalışmada 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. Solunum yolu hastalıkları, 2019 yılında dünya genelinde 4 milyon ölümle üçüncü sırada yer alarak ciddi bir küresel sağlık sorununu temsil etmiştir [5]. 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 [38, 39].

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 4 hastalık için toplamda 16.194 makale özeti elde edilmiştir. Kullanılan örnek sorgu cümleleri ve elde edilen özet sayıları Tablo 1'de gösterilmiştir.



Şekil 1. Önerilen Hastalık-Semptom İlişki Çıkarımı Yöntemi (Proposed Disease-Symptom Relation Extraction Method)

Tablo 1. Veri setine Ait Bilgiler (Information about the dataset)

Örnek Sorgu Cümleleri	Solunum Yolu Hastalığı	Özet Sayısı
coronovirus disease[Title/Abstract] AND fever[Title/Abstract] coronovirus disease[Title/Abstract] AND cough[Title/Abstract] coronovirus disease[Title/Abstract] AND loss of taste[Title/Abstract]	Covid-19	5718
bronchitis[Title/Abstract] AND sore throat [Title/Abstract] bronchitis[Title/Abstract] AND chills [Title/Abstract] bronchitis[Title/Abstract] AND hoarse voice [Title/Abstract]	Bronşit	4320
asthma[Title/Abstract] AND aphasia [Title/Abstract] asthma[Title/Abstract] AND cough[Title/Abstract] asthma[Title/Abstract] AND sleeping trouble [Title/Abstract]	Astım	1948
pulmonary embolism[Title/Abstract] AND hypotension [Title/Abstract] pulmonary embolism[Title/Abstract] AND chest pain [Title/Abstract] pulmonary embolism[Title/Abstract] AND dyspnea [Title/Abstract]	Pulmoner Emboli	4208
Toplam		*16.194*

3.2. Önerilen Semptom Çıkarım Yöntemi
(Proposed Symptom Extraction Method)

Bu bölümde, tıbbi makalelerden semptomların otomatik olarak çıkarılması amacıyla önerilen yöntem ayrıntılı bir biçimde ele alınmaktadır. Bu hedefe ulaşmak için, önceden eğitilmiş dil modeline sahip olan scispaCy [40] dil modeli, semptom ontolojisi [41] ile hibrit olarak kullanılmıştır. Böylece, scispaCy dil modelinin hastalık olarak etiketlediği semptomlar ayırt edilerek hastalık ve semptomlar arasındaki ilişkilerin tespit edilmesine olanak sağlanmıştır. Bu entegrasyon sayesinde, tıbbi makale özetlerinde yer alan semptomların otomatik olarak çıkarmak daha yüksek bir başarıda gerçekleştirilmiştir.

3.2.1. ScispaCy Dil Modeli (Scispacy Language Model)

ScispaCy, biyomedikal metin madenciliği ve doğal dil işleme alanında kullanılmak üzere özel olarak tasarlanmış bir Python kütüphanesidir. Bu kütüphane, özellikle tıbbi ve biyomedikal metinlerdeki varlık isimlerinin (örneğin hastalıklar, ilaçlar, genler) tanımlanması için geliştirilmiştir. ScispaCy, spaCy adlı popüler bir doğal dil işleme kütüphanesinin biyomedikal metinlere özgü hale getirilmiş bir versiyonudur. ScispaCy tarafından sağlanan varlık isimlerini belirlemeye yönelik dil modelleri arasında yer alan ve BC5CDR [42] veri seti üzerinde eğitilerek oluşturulan en_ner_bc5cdr_md dil modeli, metinlerdeki hastalık ve kimyasalları belirlemek için kullanılmaktadır. Ancak bu modelde hastalık ve

semptomların tümü hastalık kategorisi altında değerlendirilmektedir. Bu çalışmada en_ner_bc5cdr_md dil modeli, semptom adlarını hastalık adlarından ayırt etmek için semptom ontolojisi ile entegre edilmiştir.

3.2.2. Semptom Ontolojisi (Symptom Ontology)

Semptom ontolojisi, OBO Foundry tarafından geliştirilmiştir. OBO Foundry (Ontology for Biomedical Investigations Foundry), biyomedikal alanda kullanılan ontolojileri standardize etmeye ve yönetmeye odaklanan bir girişimdir. Bu çerçevede geliştirilen ontolojiler, bilimsel ve tıbbi araştırmaların daha tutarlı ve anlam açısından zengin bir şekilde yapılandırılmasına yardımcı olur. OBO Foundry'nin amacı, ontolojilerin standardizasyonunu sağlamak ve bilimsel araştırmalarda, veri paylaşımında ve tıbbi uygulamalarda kullanılmalarını kolaylaştırmaktır. Bu şekilde, farklı araştırmacılar, sağlık profesyonelleri ve kuruluşlar arasında bilgi paylaşımı ve iş birliği daha etkili bir şekilde gerçekleştirilebilir. Semptom ontolojisi, tıbbi alandaki semptomların, bulguların ve klinik işaretlerin yapılandırılmasını ve sınıflandırılmasını amaçlayan bir ontolojidir. Semptom ontolojisi, bir hasta tarafından bir hastalığın semptomları olarak bildirilen işlev, duyum veya görünümde algılanan değişiklikleri içeren semptomlardan oluşur ve farklı hastalıklarla ilişkilendirilen semptomları ve bu semptomların altında yatan ilişkileri tanımlar. Bu sayede, tıbbi literatürdeki semptomların daha tutarlı bir şekilde ifade edilmesi ve anlaşılması sağlanır. Semptom ontolojisi genellikle çeşitli hastalıkların semptomlarını kategorize etmek, semptomların evrimini ve gelişimini anlamak, tıbbi teşhisleri desteklemek ve tıbbi araştırmaları daha yapılandırılmış hale getirmek gibi amaçlarla kullanılır. Bu ontolojide, eş anlamlıları ve açıklamalarıyla birlikte toplam 1314 adet semptom yer almaktadır. ScispaCy ile tespit edilen her bir hastalık, semptom ontolojisindeki varlığına göre yeniden değerlendirilerek semptom veya hastalık olarak etiketlenmiştir. Örneğin, ScispaCy modelinde hastalık olarak tanımlanan diarrhoea(ishal), nausea (bulantı), vomiting (kusma) ve abdominal pain (karın ağrısı) ifadeleri semptom ontolojisinde yer aldığından semptom olarak etiketlenirken, covid-19 infection (covid-19 enfeksiyonu) hastalık olarak etiketlenmiştir. Tablo 2'de bir verilen örnek bir Pubmed makale metninden Scispacy ve önerilen semptom çıkarma yöntemi ile tespit edilen semptom ve hastalık isimlerinin karşılaştırılması yer almaktadır.

Örnek makale metni: *There was no significant difference in the incidence of **diarrhoea** (OR=1.32, 95% CI 0.8 to 2.18, Z=1.07, p=0.28, I(2)=17%) or **nausea and/or vomiting** (OR=0.96, 95% CI 0.42 to 2.19, Z=0.10, p=0.92, I(2)=55%) between either group. However, there was seven times higher odds of having **abdominal pain** in patients with severe illness when compared with non-severe patients (OR=7.17, 95% CI 1.95 to 26.34, Z=2.97, p=0.003, I(2)=0%). CONCLUSION: Our study has reiterated that GI symptoms are an important clinical feature of **COVID-19 infection**. Patients with severe disease are more likely to have **abdominal pain** as compared with patients with non-severe disease.* (Pubmed ID: 32457035)

3.3. UMLS Normalizasyonu (UMLS Normalization)

UMLS (Unified Medical Language System), tıbbi ve sağlık alanlarındaki farklı terminolojileri ve kodlamaları bir araya getirmek ve birleştirmek amacıyla geliştirilmiş bir sistemdir [43]. 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.

Tablo 2. Scispacy ve önerilen yöntem ile hastalık-semptom isimlerini çıkarma karşılaştırma sonuçları (Comparison results of extracting disease-symptom names with Scispacy and the proposed method)

Hastalık / Semptom	Etiket (Scispacy)	Etiket (Önerilen Yöntem)
Diarrhoea (ishal)	DISEASE	SYMPTOM
Nausea and/or vomiting (Bulantı ve/veya kusma)	DISEASE	SYMPTOM
Abdominal pain (karın ağrısı)	DISEASE	SYMPTOM
Covid-19 Infection (covid-19 enfeksiyonu)	DISEASE	DISEASE

Varlık bağlama, metindeki varlıkların bir bilgi tabanında karşılık gelen terimlerle otomatik olarak eşleştiren bir süreçtir [41]. Bu çalışmada bilgi tabanı olarak UMLS kullanılmıştır. UMLS'te yer alan Bağlam Benzersiz Tanımlayıcılar (Context Unique Identifiers - CUIs), 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. Tıbbi metinlerde farklı şekillerde yer alan semptomlar ve CUI karşılıklarına ait örnekler Tablo 3'te yer almaktadır.

Tablo 3. Makalelerdeki farklı semptom ifadeleri ve CUI eşleştirmeleri
(Different symptom expressions in articles and CUI matches)

Semptom	Türkçe Karşılığı	UMLS CUI
sore throat, pharyngalgia, throat pain,pain in throat	boğaz ağrısı	C0242429
headache, c)ephalgia, 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

Bu çalışmada elde edilen tüm hastalık ve semptom ifadeleri UMLS normalizasyonu uygulanmıştır. 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 ve anlam bütünlüğü sağlanmıştır.

3.4. Hastalık-Semptom İlişki Çıkarımı
(Disease-Symptom Relation Extraction)

Makale özetlerinden elde edilen semptomlar ve hastalıklar arasındaki ilişkilerin tespit edilmesi için ScispaCy tarafından sunulan, PubMed ve PMC metinleri üzerinde önceden eğitilmiş ve tıbbi alanın terimlerini etkili bir şekilde temsil eden öğrenilmiş gömme (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.

Çalışma kapsamında, üç ayrı benzerlik yöntemi kullanılmıştır. Bu yöntemler Kosinüs, Öklidyen ve nokta çarpım benzerliğidir. 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.

3.4.1. Kosinüs benzerliği (Cosinus similarity)

Kosinüs benzerliği, metin madenciliği ve doğal dil işleme gibi alanlarda sıkça kullanılan bir benzerlik ölçüsüdür. İki vektörün yönleri arasındaki açıyı ölçerek benzerlik derecesini belirler. Bu benzerlik ölçüsü, özellikle kelime vektörlerinin semantik benzerliklerini değerlendirmek için kullanılımaktadır. Kosinüs benzerliği, genellikle vektör uzayında yer alan iki vektörün iç çarpımının, vektörlerin normlarının çarpımına oranı olarak hesaplanır. İki vektör arasındaki kosinüs benzerliği, 0 ile 1 arasında bir değer alır. 1, vektörlerin tamamen örtüştüğü ve aynı yöne baktığı durumu ifade ederken, 0, vektörlerin tamamen farklı yönlere işaret ettiği anlamına gelir. Kosinüs benzerliğinin formülü Eş. 1'de verilmiştir.

$$Kosinüs\ Benzerliği = \frac{A.B}{||A||.||B||} \quad (1)$$

A ve B karşılaştırılan vektörleri, $||\boldsymbol{A}||$ ve $||\boldsymbol{B}||$ vektörlerin normlarını ifade eder.

Hastalık ve semptom vektörleri arasındaki kosinüs benzerliği, Eş. 2'de belirtildiği şekilde hesaplanır:

$$kosinusBenzerligi(V_{hastalik}, V_{semptom}) = \frac{V_{hastalik}.V_{semptom}}{||V_{hastalik}||.||V_{semptom}||} \quad (2)$$

$||V_{hastalik}||$ ve $||V_{semptom}||$ hastalık ve semptom vektör normlarını ifade eder. Eş. 3 ve Eş. 4'te belirtildiği şekilde hesaplanmıştır:

$$||V_{hastalik}|| = \sqrt{\sum_{i=1}^{n} {V_{hastalik_i}}^2} \quad (3)$$

$$||V_{semptom}|| = \sqrt{\sum_{i=1}^{n} {V_{semptom}}_i^2} \quad (4)$$

İki vektör aynı oryantasyona sahipse, aralarındaki açı 0 ve kosinüs benzerliği 1'dir. Zıt vektörlerin aralarındaki açı 180 derecedir ve kosinüs benzerliği-1'dir. Hesaplanan kosinüs benzerliği değeri, hastalık ve semptom arasındaki semantik benzerliği ölçer. Daha yüksek bir kosinüs benzerliği, hastalık ve semptomun vektör uzayında daha yakın olduğunu ve daha büyük bir semantik benzerliğe sahip olduğunu ifade eder.

3.4.2. Öklidyen benzerliği (Euclidean similarity)

Öklid benzerliği, iki vektör arasındaki benzerliği Öklidyen uzaklık ölçüsüne dayalı olarak değerlendiren bir metriktir. Öklidyen benzerliği, genellikle vektörlerin birbirine ne kadar yakın veya uzak olduğunu belirlemek için kullanılır. Öklidyen benzerliği hesaplaması, iki vektör arasındaki Öklidyen uzaklığı kullanır. Öklidyen uzaklık, iki nokta arasındaki en kısa mesafeyi temsil eder. İki vektör arasındaki Öklidyen benzerliği hesaplaması Eş. 5'te verilen formülle ifade edilir:

$$Öklidyen\ Benzerliği = \frac{1}{1+Öklidyen\ Uzaklik} \quad (5)$$

Hastalık ($V_{hastalik}$) ve semptom ($V_{semptom}$) vektörleri arasındaki öklidyen uzaklığı Eş. 6'daki formül ile hesaplanır.

$$Öklidyen\ Uzaklık = \sqrt{\sum_{i=1}^{n}(V_{hastalik_i} - V_{semptom_i})} \quad (6)$$

Öklidyen uzaklık, iki vektörün her bir bileşeninin farklarının karesinin toplamının karekökü olarak hesaplanır. Burada $V_{hastalik_i}$ ve $V_{semptom_i}$ vektörlerin i. bileşenleridir. Öklidyen benzerliği, 0 ile 1 arasında değerler alır. Daha büyük bir öklidyen benzerliği, vektörlerin daha yakın olduğunu, daha küçük bir benzerlik değeri ise vektörlerin daha uzak olduğunu ifade eder. Bu çalışmada Öklidyen benzerliği, hastalıkların ve semptomların vektör uzayında birbirine yakın veya uzak olduğunu ölçen bir metrik olarak kullanılmıştır.

3.4.3. Nokta çarpımı benzerliği (Dot product similarity)

Nokta çarpımı benzerliği (dot product similarity), iki vektörün benzerliğini ölçmek için kullanılan yöntemlerden biridir. İki vektörün nokta çarpımı benzerliği hesaplanırken, vektörlerin benzerlik düzeyi, iç çarpımları ile belirlenir. Nokta çarpımı benzerliği, iki vektörün aynı yönde ne kadar benzer olduğunu ölçer. İki vektör arasındaki açı ne kadar küçükse, nokta çarpımı benzerliği o kadar büyük olur. Eğer iki vektör tamamen aynı yönde ise, benzerlik 1'e eşit olur. Eğer vektörler tamamen zıt yönde ise, benzerlik -1'e eşit olur. Aralarındaki açı ne kadar büyükse, benzerlik değeri de o kadar küçük olur.

Hastalık ve semptom vektörleri arasındaki nokta çarpım benzerliği Eş. 7'de verilen formül ile hesaplanır.

$$Nokta\ Çarpımı\ Benzerlik = \sum_{i=1}^{n}(V_{hastalik_i} * V_{semptom_i}) \quad (7)$$

Burada $V_{hastalik}$ ve $V_{semptom}$ hastalık ve semptom vektörlerini, $V_{hastalik_i}$ ve $V_{semptom_i}$ vektörlerin i. elemanlarını, n ise vektörlerin boyutunu temsil etmektedir.

Kosinus, öklidyen ve nokta çarpımından elde edilen benzerlik puanlarına göre semptomlar hastalıklarla olan benzerlik skorlarına göre sıralanmıştır. Ters yönde kosinüs ve iç çarpım benzerlik değerleri arttıkça ilişkinin derecesi artmış, Öklidyen benzerlik değeri arttıkça ilişkinin derecesi azalmıştır. Dört solunum yolu hastalığının her biri için kosinüs, Öklidyen ve nokta çarpımı benzerlik yöntemleri kullanılarak semptom ve hastalık arasındaki benzerlik skorları hesaplandıktan sonra kosinüs ve nokta çarpımı benzerlik yöntemlerinde negatif olarak hesaplanan semptomlar semptom listesinden çıkarılarak hastalıkla ilgisiz oldukları varsayılmıştır.

3.4.4. benzerlik yöntemlerinin değerlendirilmesi (Evaluation of similarity methods)

Kosinus, öklidyen ve nokta çarpımı benzerlik yöntemleri hastalık ve semptomlar arasındaki ilişkileri tespit edebilme başarısına göre değerlendirilmiştir. Bu değerlendirme, hastalığa ait bilinen semptomların ne kadar iyi ve ne sıklıkta tespit edildiğini ölçmeye dayanmaktadır. Değerlendirme işleminde, hastalığa ait bilinen semptomların üst sıralarda tespit edilme oranı esas alınmıştır. Benzerlik yöntemlerinin başarı değerlendirmesi Eş. 8 kullanılarak hesaplanmıştır. Eşitlik 8 ile benzerlik yöntemlerinin hastalıkların bilinen semptomlarını üst sıralarda yakalayabilme oranı, dolayısıyla hastalık-semptom ilişkilerini ne kadar iyi ortaya çıkarabildiği ölçülmektedir. Hastalık bazında elde edilen semptom sayıları covid-19, bronşit, astım ve pulmoner emboli hastalıkları için sırasıyla 245, 199, 227 ve 229 adettir. Veri dağılımı göz önüne alınarak ilk 50, 100 ve 200 değerleri dikkate alınmıştır.

$$benz\_skoru = \frac{\#\,semptom50 + \#\,semptom100 + \#\,semptom200}{\#\,BilinenSemptomlar} \quad (8)$$

Bu denklemde, benz_skoru doğrulanmış semptomların ilk 50, 100 ve 200 sıralamasında görülme oranlarının toplamını,

#BilinenSemptomlar: İlgili hastalığa ait WHO (World Health Organization) ve Mayo Clinic gibi sağlık kaynaklarında doğrulanan semptom sayısını,

#semptom50: Doğrulanan semptomların ilk 50 sırada kaç kez görüldüğünü,

#symptom100: Doğrulanan semptomların ilk 100 sırada kaç kez görüldüğünü,

#symptom200: Doğrulanan semptomların ilk 200 sırada kaç kez görüldüğünü ifade eder.

Değerlendirme sonucunda elde edilen değer, benzerlik yönteminin hastalık ve semptomlar arasındaki ilişkiyi ne kadar iyi tahmin ettiğini gösterir. Daha yüksek bir değer, daha iyi bir başarıyı ifade eder. Bilinen semptomlar arasındaki hastalıkları doğru bir şekilde tespit eden benzerlik yöntemi daha yüksek bir başarı değerlendirmesi alır.

4. Deneysel Sonuçlar (Experimental Results)

Deneysel çalışmalarda, veri toplama, semptom tanımlama ve hastalık-semptom ilişkisi çıkarımı için sırasıyla Biopython, scispaCy, Scikit-learn, Numpy gibi Python kütüphaneleri kullanılmıştır. Ayrıca hastalık ve semptomlar için graf tabanlı görselleştirmeler Networkx kütüphanesi ile oluşturulmuştur. Deneysel çalışmalar Google Colab ortamında, GPU desteği ile gerçekleştirilmiştir.

4.1. Semptom Çıkarımı Sonuçları (Symptom Extraction Results)

Dört solunum yolu hastalığına ait makale özetlerinden Scispacy dil modeli (en_ner_bc5cdr_md) ve semptom ontolojisi kullanılarak her bir hastalığa ait özetler içerisinden ilgili hastalık ile ilişkili olabilecek semptomlar elde edilmiştir. Makale özetlerinde belirtilen semptomların hastalıkla doğrudan veya dolaylı olarak ilişkili olduğu varsayılmaktadır. Önerilen semptom çıkarım yöntemi kullanılarak, toplam 16.194 makale özetinden toplam 1138 adet semptom elde edilmiştir. UMLS filtreleme işlemleri sonucunda, dört hastalık için tanımlanan toplam semptom sayısı 900 olarak belirlenmiştir. Her bir hastalık için birçok ortak semptom olduğundan tespit edilen benzersiz semptom sayısı 304 adettir. Çıkarılan semptom sayılarına ait bilgiler Tablo 4'te verilmiştir.

Tablo 4. Çıkarılan Semptom Sayıları (Number of extracted symptoms)

Hastalık	Semptom Sayısı	Normalizasyon Sonrası Semptom Sayısı
Covid-19	325	245
Bronşit	248	199
Astım	285	227
Pulmoner Emboli	280	229
Toplam	*1138*	*900*

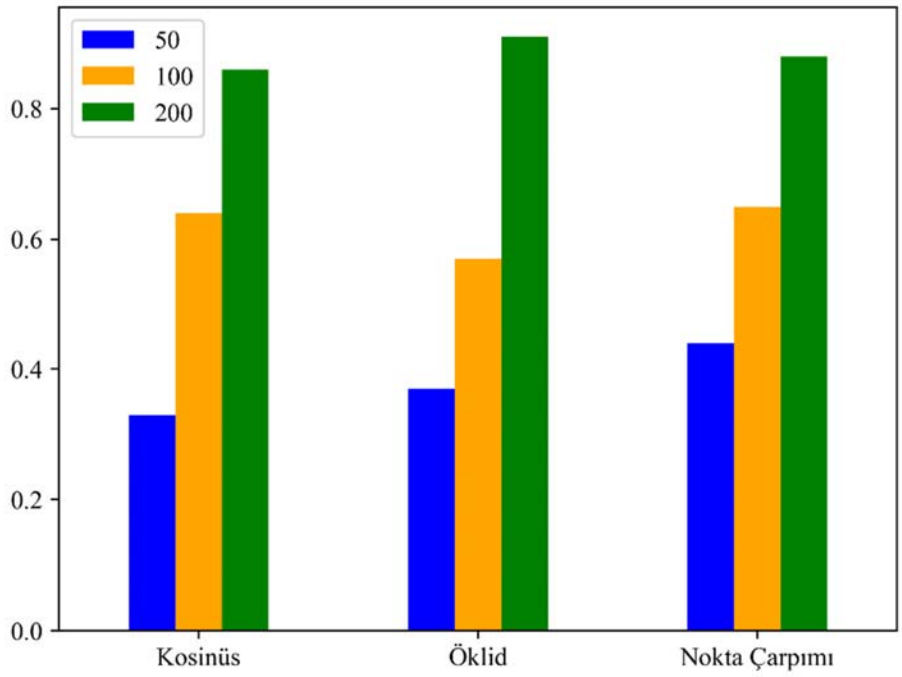
4.2. İlişki Çıkarımı Sonuçları (Relation Extraction Results)

Solunum yolu hastalıklarına ait vektörler ile elde edilen semptomların vektörel mesafesi üç farklı benzerlik yöntemi ile ölçülerek her bir hastalık-semptom çifti için üçer adet ilişki skoru hesaplanmıştır. Elde edilen skorlar temel alınarak semptomlar hastalıklara semantik uzaklığına göre sıralanmıştır. Benzerlik yöntemlerinin değerlendirilmesinde ise semantik uzaklık sıralamalardan faydalanılmıştır. Tablo 5'te her bir benzerlik yöntemi için hastalık bazında #semptom50, #semptom100, #semptom200 ve #BilinenSemptomlar değişkenlerine ait değerler verilmiştir. Bu değerler ile her bir hastalık için benzerlik yöntemlerinin benzerlik başarı skorları Eş. 2, Eş. 6 ve Eş. 7 kullanılarak hesaplanmış, yöntem bazında ortalama değerler elde edilmiştir. Kosinüs, öklid ve nokta çarpım benzerliği yöntemlerinin ortalama değerlendirme puanları sırasıyla 0,61, 0,62 ve 0,66'dır. Şekil 2'de hastalıklar için benzerlik yöntemlerinin değerlendirme puanlarını gösterilmiştir. Nokta çarpım benzerliği yöntemi, 0,66'lık ortalama benz_skoru ile diğer yöntemlerden daha iyi performans göstermiştir. Nokta çarpımı benzerliği yönteminin diğer yöntemlere göre daha yüksek benz_skoru'na sahip olması, bu yöntemin hastalık semptomlarını daha iyi tespit etme yeteneğine işaret etmektedir. Bu yöntem hastalıkla ilişkili semptomları daha üst sıralarda yerleştirme konusunda daha başarılıdır.

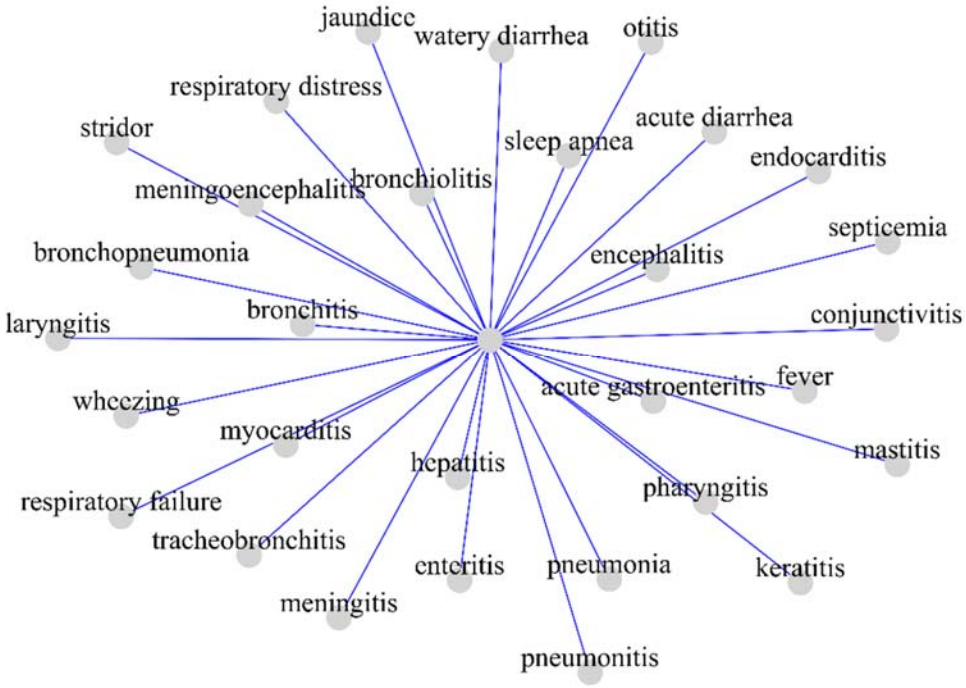
Hastalık-Semptom ilişkilerini görselleştirmek için düğüm-bağlantı grafikleri oldukça etkili bir araçtır. Bu tür grafikler, hastalıklar, semptomlar ve bunlar arasındaki ilişkileri daha karmaşık bir şekilde ifade edebilirler. Hastalıklar ve semptomlar düğümler (nodes) olarak

Tablo 5. Benzerlik Yöntemleri İçin Değerlendirme Sonuçları (Evaluation Results for Similarity Methods)

Benzerlik Yöntemi	Hastalık	Sağlık Kaynaklarından 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

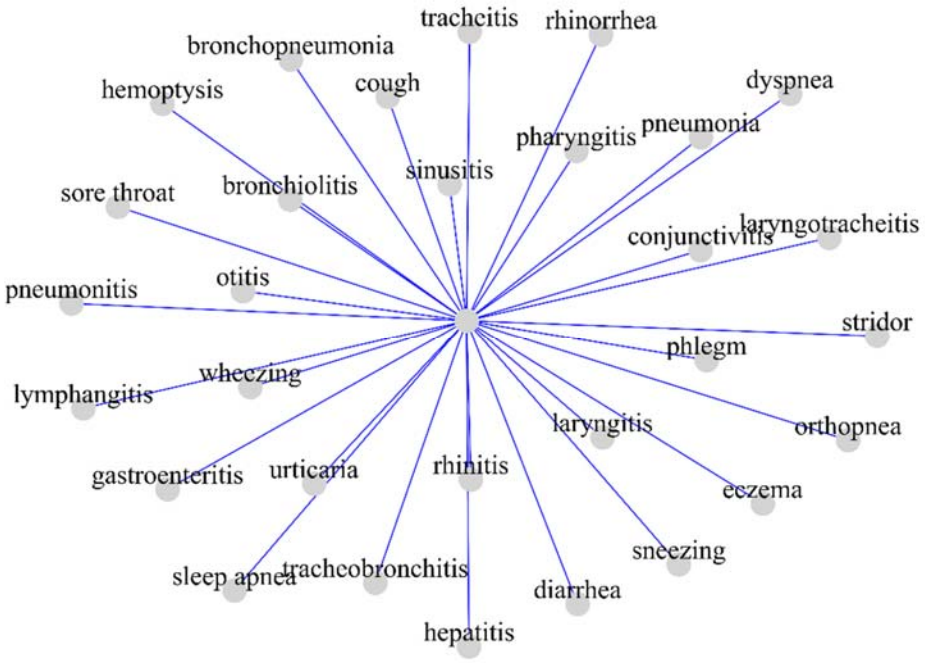


Şekil 2. Benzerlik Yöntemleri İçin Değerlendirme Sonuçlarının Ortalaması (Average of Evaluation Results for Similarity Methods)

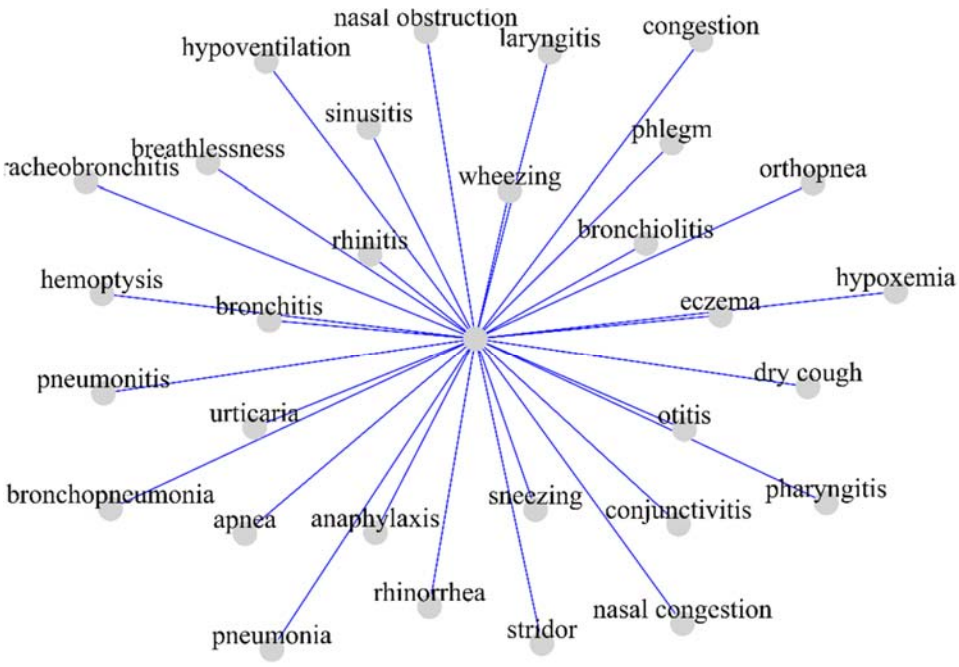


Şekil 3. Covid-19 ile İlişki Derecesi En Yüksek 30 Semptom (30 Symptoms Most Associated with Covid-19)

temsil edilir. Hastalıklarla semptomlar arasındaki ilişkiler ise bağlantılar (links) ile gösterilir. Örneğin, bir hastalık düğümü ile bu hastalığın en yaygın semptomlarını bağlayan bağlantılar kullanarak hastalık-semptom ilişkisini görsel olarak gösterilebilir. Şekil 3, Şekil 4, Şekil 5 ve Şekil 6'da Covid-19, bronşit, astım ve pulmoner emboli hastalıkları ile nokta çarpımı benzerlik yöntemine göre ilişki derecesi en yüksek 50 semptom görselleştirilmiştir. Görselleştirme işlemleri için Networkx kütüphanesi kullanılarak hastalık ve semptomların düğüm (node), ve semptomların ilgili hastalığa benzerlik derecesinin bağlantılar (edges) olarak temsil eden bir graf oluşturulmuştur. Benzerlik puanları, bağlantıların kalınlığını ve semptomun hastalığa olan uzaklığını belirlemek için kullanılmıştır. Benzerlik puanları negatif olan semptomlar listeden çıkarıldıktan sonra, geriye kalan benzerlik listesindeki son 10 adet semptomun ilgili hastalık ile ilişkisi



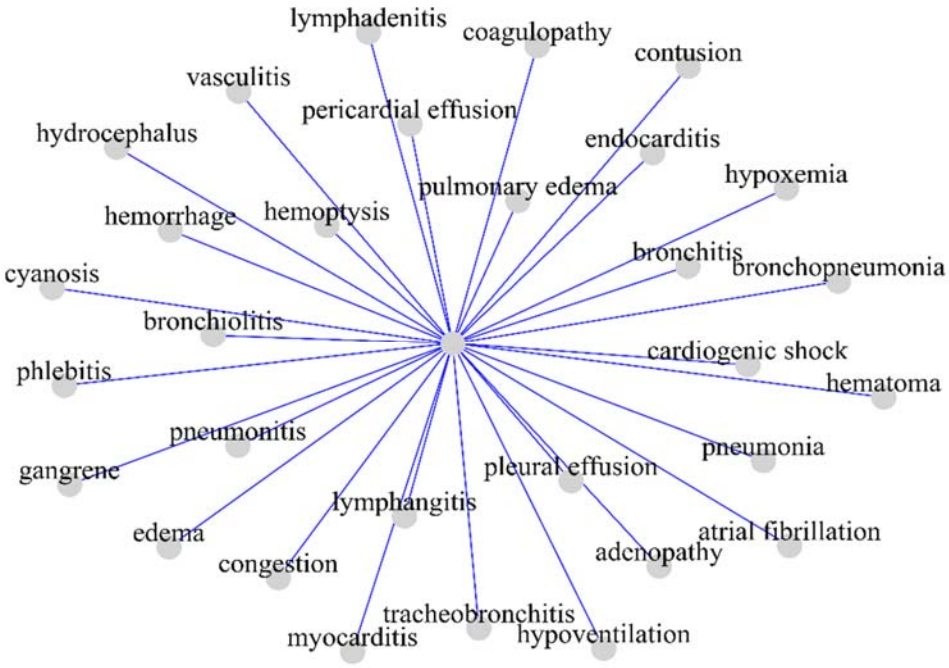
Şekil 4. Bronşit ile İlişki Derecesi En Yüksek 30 Semptom (30 Symptoms Most Associated with Bronchities)



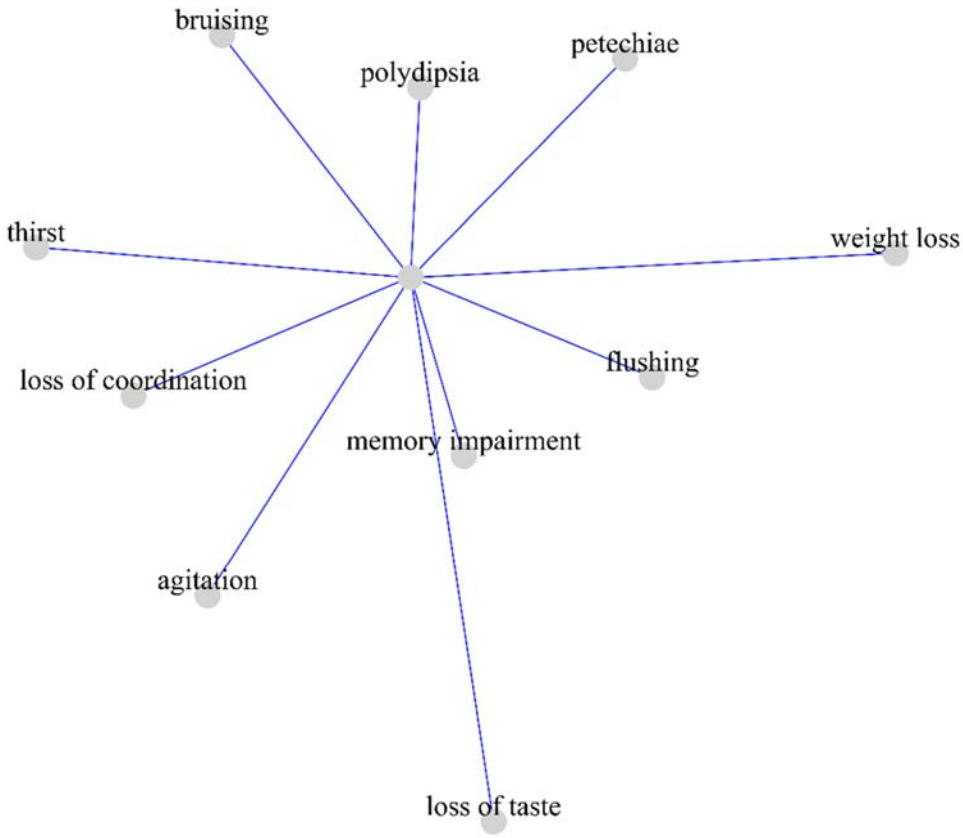
Şekil 5. Astım ile İlişki Derecesi En Yüksek 30 Semptom (30 Symptoms Most Associated with Asthma)

araştırılmıştır. Yapılan literatür araştırmalarında nadir olarak belirlenen düşük skorlu semptomların ilgili hastalıklar ile benzerlik skorları düşük bile olsa bu semptomların ilişkili olduğuna dair bulgular yer almaktadır. Yapılan literatür doğrulaması bu semptomların nadir olsalar da belirli hastalıklar ile ilişkili olabileceğini ve tıbbi teşhis süreçlerinde dikkate alınması gerektiğini göstermektedir. Örneğin, Covid-19 hastalığının hafıza kaybı (memory loss) ile [44], bronşit hastalığının deri lezyonu ile [45], astım hastalığının agresif davranış ile [46] ilişkili olduğuna dair klinik çalışmalar mevcuttur . Hastalıklar ve nadir semptomlarına ait ilişki grafları Şekil 7, Şekil 8, Şekil 9 ve Şekil 10'da gösterilmiştir. Solunum yolu hastalığı veri setinde önerilen semptom çıkarma yönteminin ve hastalık-semptom ilişkisi çıkarma mimarisinin deneysel sonuçları birkaç fikir sağlamıştır:

- Önerilen semptom çıkarma yöntemimiz, scispaCy biyomedikal varlık ismi tanıma modeline kıyasla umut verici sonuçlar vermiştir. scispaCy en_ner_bc5cdr_md modeli hastalık ve semptomları ayırt



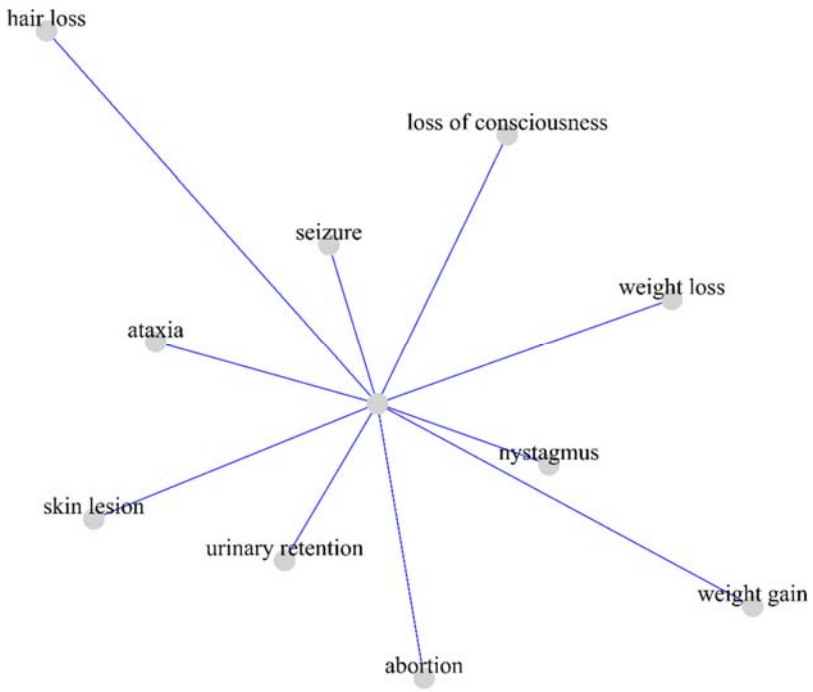
Şekil 6. Pulmoner Emboli ile İlişki Derecesi En Yüksek 30 Semptom (30 Symptoms Most Associated with Pulmonary Embolism)



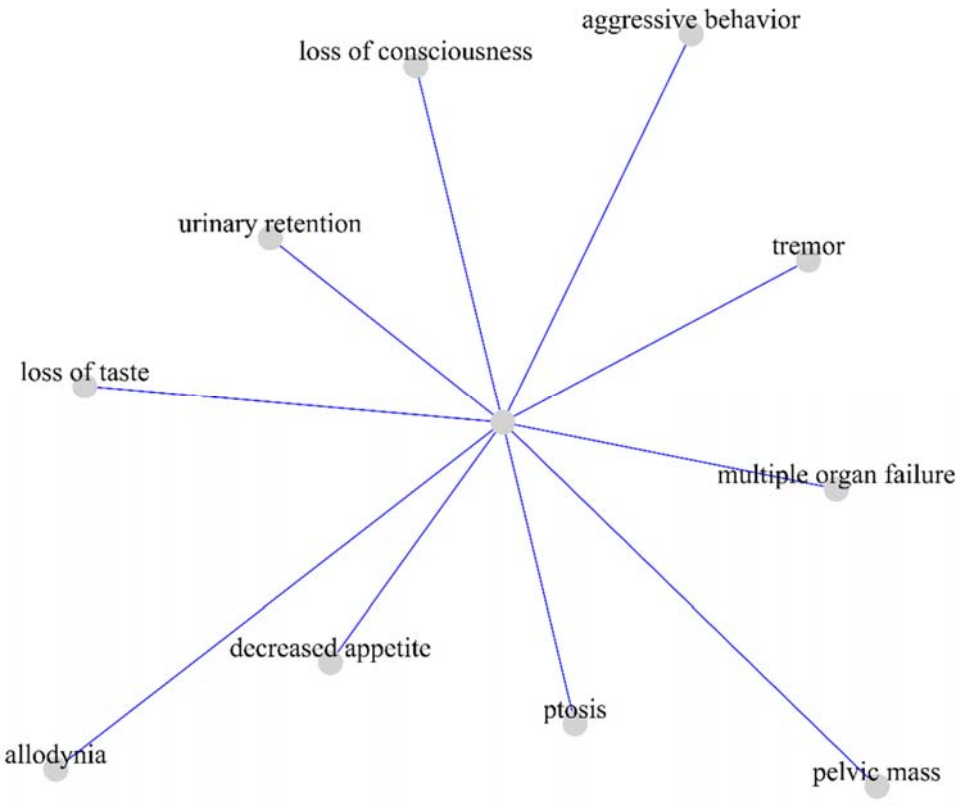
Şekil 7. Covid-19 ile İlişki Derecesi En Düşük olan 10 Semptom (10 Symptoms Least Linked to Covid-19)

edemezken, bu model semptom ontolojisi ile zenginleştirildiğinde tıbbi metinlerdeki semptomları başarılı bir şekilde tespit edebilmiştir.

- Nokta çarpım benzerlik yöntemi, özellikle üst sıralarda ilgili solunum semptomlarını tahmin etmede diğer benzerlik yöntemlerinden daha iyi performans göstermiştir.

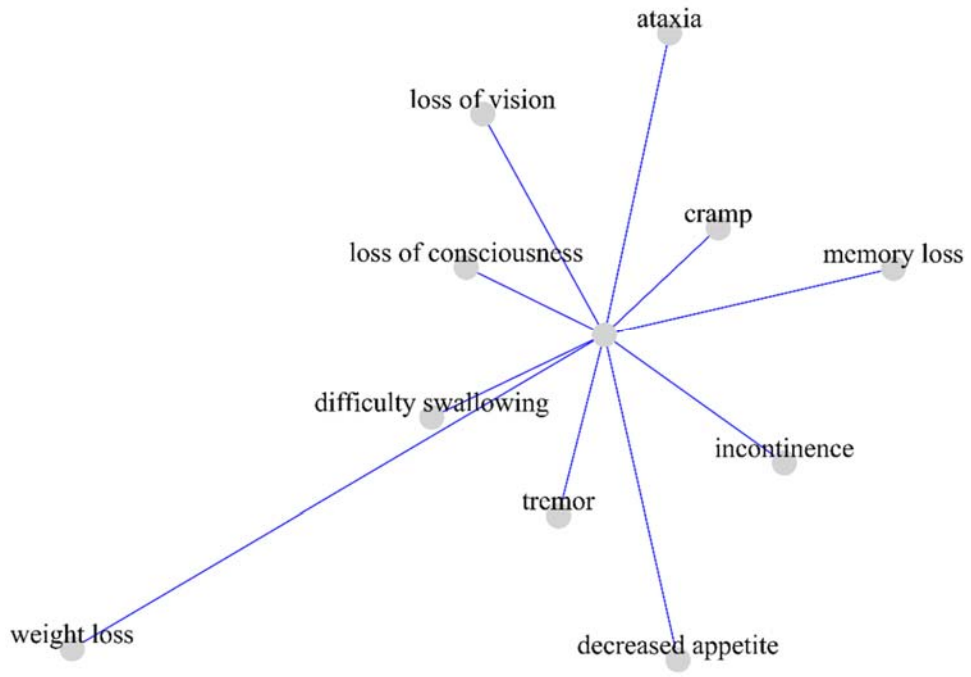


Şekil 8. Bronşit ile İlişki Derecesi En Düşük olan 10 Semptom (10 Symptoms Least Linked to Bronchities)



Şekil 9. Astım ile İlişki Derecesi En Düşük olan 10 Semptom (10 Symptoms Least Linked to Asthma)

- Hastalık-semptom ilişkisi çıkarma yöntemi, hastalıkla ilişkili olması muhtemel olan ancak biyomedikal literatürde bahsedilmeyen nadir semptomları belirlemek için çok uygundur.
- Bilimsel kanıtlar, bazı nadir semptomlar ile bunlara karşılık gelen hastalıklar arasındaki ilişkiyi desteklemektedir. Örneğin, susuzluk diyabetin bir semptomu olmasına rağmen Covid-19 ile ilgili makalelerde de bahsedilmektedir, bu da susuzluk ile Covid-19 arasında dolaylı bir ilişkiye işaret edebilir. Benzer şekilde, kramplar, uzun uçak yolculuğu sırasında uzun süreli, sıkışık oturmadan kaynaklanan pulmoner emboli ile ilişkilendirilmiştir.



Şekil 10. Pulmoner Emboli ile İlişki Derecesi En Düşük olan 10 Semptom (10 Symptoms Least Linked to Pulmonary Embolism)

Yüksek C-reaktif protein seviyeleri, bozulmuş böbrek fonksiyonu ve uzayan Covid-19 süresi de kilo kaybı ile ilişkilendirilmiştir.

- Sonuçlar, metin madenciliği yöntemlerinin solunum yolu hastalıkları gibi benzer hastalıklarının ilişkili ve nadir semptomlarını çıkarma konusunda potansiyel olarak doktorlara yardımcı olabileceğini göstermektedir.

5. Sonuçlar (Conclusions)

Bu çalışmada, biyomedikal bilgi çıkarımının bir alt alanı olan hastalık semptom ilişki çıkarımı problemine yönelik bir çözüm önerisi sunulmuştur. Semptom ve hastalık isimlerinin biyomedikal literatürden tespit edilmesi için hibrit bir varlık ismi tanıma yöntemi ve ilişki çıkarımı için semantik benzerlik yöntemlerine dayalı bir çözüm önerilmiştir. Solunum yolu hastalıkları ve semptomları arasındaki ilişkileri çıkarmak için kosinüs, öklidyen ve nokta çarpımı benzerlik yöntemlerine dayalı yaklaşımlar ile hastalıklar ve semptomlar arasındaki ilişki dereceleri tespit edilmiştir. Bu yöntemlerin sonuçları değerlendirildiğinde nokta çarpımı benzerlik yönteminin 0,66 ortalama benz_skoru hastalık ve semptom ilişkisini tespit etmede diğer semantik benzerlik yöntemlerine göre daha yüksek bir başarıya sahip olduğu görülmüştür. Nokta çarpımı benzerliği, WHO ve Mayo Clinic gibi sağlık kaynaklarında belirtilen semptomların ilk sıralarda tespit edilebilmektedir. İlişki analizindeki bir diğer amacımız sağlık kaynaklarında doğrudan belirtilmeyen ancak biyomedikal literatürde gizlenen semptomları keşfetmekti. Hastalıklar ile ilişki (benzerlik) derecesi düşük olan ancak potansiyel olarak hastalığın bir göstergesi olabilecek semptomlar nadir semptom olarak belirtilmiş ve bu semptomların hastalıklar ile ilişkili olabileceği literatürden bilimsel çalışmalar örnek gösterilerek doğrulaması gerçekleştirilmiştir. Elde edilen sonuçlar, hekimlere hastalıklarla ilgili şüpheli olabilecek hususlarda, az görülen semptomların da gözden kaçırılmadan değerlendirilmesini sağlayarak tanı ve teşhis süreçlerine katkı sağlayabilir. Covid-19 gibi artan salgın hastalık varyantları ile semptomlar değişim geçirebilmesi de göz önüne alındığında güncel semptomların ve hastalıkların manuel olarak kısıtlı zaman içerisinde takip edilmesi, incelenmesi ve analiz edilmesi oldukça zordur. Önerilen yöntem ile, ekstra zaman kaybı ve çaba gerekmeden hekimlerin hastalık ve semptomlar ile ilgili güncel gelişmelere ulaşabilme ve analiz edebilmesine katkı sağlayabilir. Bu çalışma, solunum yolu hastalıklarına odaklanmıştır, ancak aynı yöntemler diğer hastalık grupları için de uygulanabilir. Gelecek çalışmalarda, farklı hastalık türleri üzerinde hastalık-semptom ilişkilerini çıkarmayı hedeflenmektedir. Daha büyük ve çeşitli biyomedikal metin veri kümeleri kullanarak yöntemlerin genellenebilirliği ve performansı daha ayrıntılı bir şekilde değerlendirilebilir.

Kaynaklar (References)

1. Eurostat Statistics Explained, Respiratory diseases statistics, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Respiratory_diseases_statistics&oldid=541149#Deaths_from_diseases_of_the_respiratory_system,Erişim Tarihi Mart 10 2023.
2. Brosnahan, S. B., Jonkman, A. H., Kugler, M. C., Munger, J. S., & Kaufman, D. A. (COVID-19 and Respiratory System Disorders: Current Knowledge, Future Clinical and Translational Research Questions. Arteriosclerosis, Thrombosis, and Vascular Biology, 40 (11), 2586-2597, 2020.
3. PUBMED, pubmed.gov.tr. Erişim Tarihi Aralık 10 2023.
4. PubMed Overview ,https://pubmed.ncbi.nlm.nih.gov/about/#:~:text=PubMed%20Overview ,health%E2%80%93both%20globally%20and%20personally, Erişim Tarihi Aralık 10 2023.
5. MEDLINE PubMed Production Statistics, https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html, Erişim Tarihi Aralık 10 2023.
6. Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., & Wang, J. Attention guided capsule networks for chemical-protein interaction extraction. Journal of Biomedical Informatics, 103, 103392, 2020.
7. Peng, Y., Rios, A., Kavuluru, R., & Lu, Z. Extracting chemical–protein relations with ensembles of SVM and deep learning models. Database, 2018.
8. Zhou, H., Liu, Z., Ning, S., Lang, C., Lin, Y., & Du, L. Knowledge-aware attention network for protein-protein interaction extraction. Journal of Biomedical Informatics, 96, 103234, 2019.

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

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

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

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

13. Machado, J., Rodrigues, C., Sousa, R., & Gomes, L. M. Drug–drug interaction extraction-based system: An natural language processing approach. Expert Systems, e13303, 2023.

14. Li, J., Sun, A., Han, J., & Li, C. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34 (1), 50-70, 2020.

15. Yang Z, Lin H, Li Y. Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature. Comput Biol Chem, 32 (4), 287–91, 2008.

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

17. N Kang, B Singh, Z Afzal, et al. Using rule-based natural language processing to improve disease normalization in biomedical text. J Am Med Inform Assoc. 20 (5), 876–81, 2013.

18. Fukuda, K. I., Tsunoda, T., Tamura, A., & Takagi, T. Toward information extraction: identifying protein names from biological papers. In Pac symp biocomput, 707 (18), 707-718, 1998.

19. İlgün E.G., Samet R., Increasing the performance of intrusion detection models developed using machine learning method with preprocessing applied to the dataset Journal of the Faculty of Engineering and Architecture of Gazi University, 39 (2), 679-692, 2024.

20. Morwal, S., Jahan, N., & Chopra, D. Named entity recognition using hidden Markov model (HMM). International Journal on Natural Language Computing (IJNLC), 1, 2012.

21. Xie, X. Y. A Review on Support Vector Machines for Biomedical NER. Data Science for NLP, 1, 06, 2020.

22. Cho, M., Ha, J., Park, C., & Park, S. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. Journal of biomedical informatics, 103, 103381, 2020.

23. Zhu, Q., Li, X., Conesa, A., & Pereira, C. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. Bioinformatics, 34 (9), 1547-1554, 2018.

24. Jackson RG, Patel R, Jayatilleke N, et al Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) Project BMJ Open;7:e012012. doi: 10.1136/bmjopen-2016-012012, 2017.

25. Wu, C. S., Kuo, C. J., Su, C. H., Wang, S. H., & Dai, H. J. 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, 2020.

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

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

28. Leiter, R. E., Santus, E., Jin, Z., Lee, K. C., Yusufov, M., Chien, I., ... & Lindvall, C. 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., 2020.

29. Wang, J., Abu-el-Rub, N., Gray, J., Pham, H. A., Zhou, Y., Manion, F. J., ... & Zhang, Y. 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, 2021.

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

31. Du, N., Chen, K., Kannan, A., Tran, L., Chen, Y., & Shafran, I., Extracting symptoms and their status from clinical conversations. arXiv preprint arXiv:1906.02239, 2019.

32. Alshuwaier, F., Areshey, A., & Poon, J. A comparative study of the current technologies and approaches of relation extraction in biomedical literature using text mining. In 2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS), 1-13, IEEE, 2017.

33. Zhou X, Menche J, Barabási A-L, et al. Human symptoms–disease network. Nat Commun. 5, 4212, 2014.

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

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

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

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

38. Ma X., Conrad T., Alchikh M., Reiche J., Schweiger B., Rath B.. 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, 2018.

39. Van der Sar IG., Wijsenbeek MS., Braunstahl GJ., Loekabino JO., Dingemans AC., In 't Veen JCCM, Moor CC. Differentiating interstitial lung diseases from other respiratory diseases using electronic nose technology. Respir Res. 24 (1), 271, 2023.

40. Neumann, M., King, D., Beltagy, I., & Ammar, W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. ArXiv, abs/1902.07669., 2019.

41. The Open Biological and Biomedical Ontology (OBO) Foundry, "Symptom Ontology", https://obofoundry.org/ontology/symp.html, Erişim 03.12.2021.

42. Jiao L., Yueping S., Robin J., Daniela S., Chih-Hsuan W., Robert L, Allan P.D., Carolyn J.M., Thomas C.W., and Zhiyong L., BioCreative V CDR task corpus: a resource for chemical disease relation extraction, Database: the journal of biological databases and curation, 2016.

43. Rao, D., McNamee, P., & Dredze, M. Entity linking: Finding extracted entities in a knowledge base, Multi-source, multilingual information extraction and summarization, Theory and Applications of Natural Language Processing. Springer, Berlin, Heidelberg 93-115, 2013.

44. Ahmed M., Roy S., Iktidar MA., Chowdhury S., Akhter S., Khairul Islam AM., Hawlader MDH. Post-COVID-19 Memory Complaints: Prevalence and Associated Factors, Neurologia. PMID: 35469238; PMCID: PMC9020525, 2022.

45. Shah PL., Orton C. Epithelial Resurfacing: The Bronchial Skin Peel. American Journal of Respiratory and Critical Care Medicine, 202 (5), 641-642, 2020.

46. Lehrer, P. M., Isenberg, S., & Hochron, S. M. Asthma and emotion: a review. Journal of Asthma, 30 (1), 5-21, 1993.