T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Yüksek Lisans Tezi

RADYOLOJİ ALANINDA MAKİNE ÖĞRENME YÖNTEMLERİNİN KULLANILDIĞI YAYINLARIN DOĞAL DİL İŞLEME İLE ANALİZİ

Selin ŞAHİN

Enformatik Anabilim Dalı

Enformatik Programı

DANIŞMAN
Prof. Dr. Sevinç GÜLSEÇEN

II. DANIŞMAN
Dr. Serra ÇELİK

Mayıs, 2023

İSTANBUL

Bu çalışma, 4.05.2016 tarihinde aşağıdaki jüri tarafından Enformatik Anabilim Dalı, Enformatik Programında Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Prof. Dr. Sevinç GÜLSEÇEN(Danışman)
İstanbul Üniversitesi
Enformatik Bölümü

Doç. Dr. Çiğdem EROL
İstanbul Üniversitesi
Enformatik Bölümü

Doç. Dr. M. Fevzi ESEN
Sağlık Bilimleri Üniversitesi
Hamidiye Sağlık Bilimleri Enstitüsü

İntihal Programı Beyanı

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

Proje Destekleri

Bu tez, İstanbul Üniversitesi Bilimsel Araştırma Projeleri Yürütücü Sekreterliğinin ..... numaralı projesi ile desteklenmiştir.

Bu tez, ... numaralı ..... projesi ile desteklenmiştir.

Tezden Üretilmiş Yayınların Künye Bilgileri

ÖNSÖZ

"Radyoloji Alanında Makine Öğrenme Yöntemlerinin Kullanıldığı Yayınların Doğal Dil İşleme ile Analizi" isimli radyolojide makine öğrenmesi yöntemlerinden yararlanılan makale çalışmalarının özetleri üzerinden doğal dil işleme yöntemleri kullanılarak bir bibliyografik analiz gerçekleştirdiğim çalışmamın her aşamasında bana rehberlik eden tezime zaman ayırıp tavsiyeler sunan kıymetli hocam Prof. Dr. Sevinç GÜLSEÇEN ve sonsuz destek ve yardımını gördüğüm kıymetli hocam Dr. Serra ÇELİK'e teşekkürlerimi sunarım.

Her zaman destekçim olan canım annem Güllü ŞAHİN ve canım babam Murat ŞAHİN'e, motivasyonumu yükselten ve bana destek olan tek tek isimlerini sayamayacağım arkadaşlarıma sonsuz teşekkürlerimi sunarım.

6 Şubat 2023 tarihinde meydana gelen depremden etkilenen Hatay'da kaybettiğim ve akademiye olan ilgisiyle beni de etkileyen sevgili kuzenim Gülşah ŞAHİN AKI'yı sevgi ve rahmetle anmayı bir borç bilirim.

Ayrıca tez çalışmasında analizler için gerekli bilgisayar gücünü sağladığı için İstanbul Üniversitesi Gerçek Veri Uygulama ve Araştırma Laboratuvarı'na teşekkür ederim.

Mayıs 2023 Selin ŞAHİN

İÇİNDEKİLER

ÖNSÖZ ..... **iv**
İÇİNDEKİLER ..... **v**
ŞEKİL LİSTESİ ..... **vii**
TABLO LİSTESİ ..... **viii**
KISALTMA LİSTESİ ..... **ix**
ÖZET ..... **x**
SUMMARY ..... **xii**
1.GİRİŞ **1**
1.1 DOĞAL DİL İŞLEME 1
1.1.1 Doğal Dil İşleme Uygulama Alanları 2
1.2 RADYOLOJİ VE DOĞAL DİL İŞLEME 4
2.GENEL KISIMLAR **7**
2.1 BİBLİYOGRAFİK ÇALIŞMA 7
2.2 BİLGİ ÇIKARIMI 7
2.2.1 Adlandırılmış Varlık Tanıma (Named Entity Recognition (NER)) 8
2.2.2 Sözdizimsel Analiz (Syntactic Analysis) 9
2.2.3 Anlamsal Analiz (Semantic Analysis) 9
2.2.4 Eş referans Çözünürlüğü (Co-reference Resolution (CO)) 9
2.2.5 İlişki Çıkarımı (Relation Extraction (RE)) 9
2.2.6 Olay Çıkarma (Event Extraction (EE)) 9
2.3 KONU MODELLEME 10
2.3.1 Gizli Anlamsal Analiz 11
2.3.2 Negatif Olmayan Matris Çarpanlarına Ayırma 11
2.3.3 Gizli Dirichlet Ayrımı 12
2.3.4 Pachinko Dağılım Modeli 13
3. MALZEME VE YÖNTEM **16**
3.1 VERİ SETİ 16
3.2 VERİ ÖNİŞLEME 17
3.3 VERİ ANALİZ 19
3.3.1 Veri Temizleme ve Ön İşleme 19
4. BULGULAR **29**
5. TARTIŞMA VE SONUÇ **31**

KAYNAKLAR **33**
ÖZGEÇMİŞ **40**

ŞEKİL LİSTESİ

Sayfa No

Şekil 1.1: Radyolojide Doğal Dil İşleme (Türkçeye uyarlanmıştır) (Cai ve diğ, 2016).....6

Şekil 3.1 Veri Çekmede Kullanılan Alan Kısaltmaları 17

Şekil 3.2: Çalışma Metodolojisi (Guan ve diğ, 2019) 18

Şekil 3.3 Veri Önişleme Aşamasında Kaldırılan Semboller 19

Şekil 3.4 Metinde En Fazla Kullanılan Kelimeler ile Histogram Grafiği (durak kelimeler ile).....20

Şekil 3.5 Metinde En Fazla Kullanılan Kelimeler ile Histogram Grafiği (durak kelimeler kaldırılarak) 20

Şekil 3.6 Önişleme Sonucu Elde Edilen Özet Metinler.....21

Şekil 3.7 En sık kullanılan kelimeler ile kelime bulutu (Word Cloud) 21

Şekil 3.8 Coherence Score Grafiği 22

Şekil 3.9 Konu 1 için LDA tarafından tanımlanan konular için pyLDAvis grafiği 23

Şekil 3.10 Konu 1'e ait topic örneği.....24

Şekil 3.11 Konu 2 için LDA tarafından tanımlanan konular için pyLDAvis grafiği 24

Şekil 3.12 Konu 3 için LDA tarafından tanımlanan konular için pyLDAvis grafiği 25

Şekil 3.13 Konu 3'e ait topic örneği.....25

Şekil 3.14 Konu 4 için LDA tarafından tanımlanan konular için pyLDAvis grafiği 26

Şekil 3.15 Konu 4'e ait topic örneği.....26

Şekil 3.16 Konu 5 için LDA tarafından tanımlanan konular için pyLDAvis grafiği 27

Şekil 3.17 Konu 5'e ait topic örneği.....27

Şekil 4.1 PubMed'de Yıllara Göre Yayımlanan Çalışma Sayıları (Veri Önişleme Yapıldıktan Sonra Elde Edilen Özetlerin Dağılımı).....29

TABLO LİSTESİ

Sayfa No

Tablo 3.1 Konu Tutarlılık Skoru 22

Tablo 4.1 PubMed'de Yıllara Göre Yayımlanan Çalışma Sayıları.....29

KISALTMA LİSTESİ

Kısaltmalar **Açıklama**

BoW : Bag of Words

CALL : Computer Assisted Language Learning

CO : Co-referance Resolution

CT : Computerized Tomography

DDİ : Doğal Dil İşleme

EE : Event Extraction

IDF : Inverse Document Frequency

IE : Information Extraction

IR : Information Retrieval

LDA : Latent Dirichlet Allocation

LSA : Latent Semantic Analysis

MRI : Magnetic Resonance Imaging

NER : Named Entity Recognition

NMF : NonNegative-Matrix Factorization

PAM : Pachinko Allocation Model

PET : Positron Emission Tomography

RE : Relation Extraction

QA : Question Answering

TF : Term Frequency

ÖZET

YÜKSEK LİSANS TEZİ

RADYOLOJİ ALANINDA MAKİNE ÖĞRENME YÖNTEMLERİNİN KULLANILDIĞI YAYINLARIN DOĞAL DİL İŞLEME İLE ANALİZİ

Selin ŞAHİN

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Enformatik Anabilim Dalı

Danışman : Prof. Dr. Sevinç GÜLSEÇEN

II. Danışman : Dr. Serra ÇELİK

İnsan dilinin makine aracılığı ile işlenmesi olarak bilinen Doğal Dil İşleme kavramı gün geçtikçe katlanarak artan metin verilerinden öngörü çıkarılması amacıyla daha sık karşımıza çıkmaktadır. Tez çalışmasında 2017-2023 yılları arasında radyoloji alanında makine öğrenmesi yöntemleri kullanılarak yapılan çalışmalar PubMed üzerinden çekilerek 9582 makale özetini kapsayan bir veri tabanı oluşturulmuş ve bunların bibliyografik analizi gerçekleştirilmiştir. Analizde Pyhton programlama dili kütüphanelerinden yararlanılmıştır. Analiz sonucunda makale özetleri tıbbi hastalıklar, radiomics, kanser ve görüntüleme, klinik yapay zeka ve makine öğrenmesi (sınıflandırma) olmak üzere 5 ana konu başlığı altında toplanmıştır. Latent Dirichlet Allocation (LDA) yöntemi konu modelleme için sıklıkla kullanılan bir yöntem olmakla birlikte özellikle kelimeler arası bağlılıkları ortaya çıkarmada bazı durumlarda yetersiz kalabilmektedir. Bundan dolayı n-gram gibi kelimeler arası ilişki kuran ileri doğal dil işleme yöntemlerinin kullanılması gerekebilir.

Mayıs 2023, 54 sayfa.

Anahtar kelimeler: lda, konu modelleme, doğal dil işleme

SUMMARY

M.Sc. THESIS

NATURAL LANGUAGE PROCESSING ANALYSIS OF PUBLICATIONS IN RADIOLOGY USING MACHINE LEARNING METHODS

Selin ŞAHİN

İstanbul University

Institute of Graduate Studies in Sciences

Department of Informatics

Supervisor : Prof. Dr. Sevinç GÜLSEÇEN
Co-Supervisor : Dr. Serra ÇELİK

The concept of Natural Language Processing, known as the processing of human language by machine, appears more frequently in order to extract predictions from text data that is increasing exponentially day by day.In the thesis study, a database was created by collecting the articles carried out using machine learning methods in the field of radiology between the years 2017-2022 via PubMed.And a bibliographic analysis was made with these article summaries. These database contains 9582 article summaries. Python programming language was used in the analysis. As a result of the analysis, the abstracts of the articles were clustered around 5 main topics: medical diseases, radiomics, cancer and imaging, clinical artificial intelligence and machine learning (classification). Although the Latent Dirichlet Allocation (LDA) method is a frequently used method for topic modeling, it may be insufficient in some cases, especially in revealing inter-word dependencies. Therefore, it may be necessary to use advanced natural language processing methods that establish relationships between words such as n-gram.

May 2023, 54 pages.

Keywords: lda, topic modeling, natural language processing

1.GİRİŞ

Radyoloji alanında son yıllarda görüntü işleme başta olmak üzere makine öğrenmesi yöntemlerine başvurulduğu görülmektedir. Bu nedenle bu tez çalışmasının konusunu radyoloji ve makine öğrenmesi oluşturmuş ve trendin ne yönde geliştiği incelenmiştir.

Çalışmanın birinci bölümünde Doğal Dil İşleme kavramı, doğal dil işleme ile bağlantılı olan kavramlar ile doğal dil işlemenin uygulama alanları ele alınmış ayrıca çalışma genelinde neden radyoloji ile ilgilenildiği açıklanmıştır.

İkinci bölümde bilgi çıkarımı kavramı, bilgi çıkarımı için kullanılan analiz türleri, konu modelleme ve konu modelleme kapsamında kullanılan yöntemler yer almıştır.

Üçüncü bölümde veri seti tanıtılarak veri ön işleme ile ilgili genel bilgiler verilmiştir.

Son bölüm olan dördüncü bölümde ise veri analizi gerçekleştirilerek sonuçları tartışılmıştır.

1.1 DOĞAL DİL İŞLEME

Doğal diller, kendiliğinden ortaya çıkan ve insanlar tarafından iletişim için kullanılan dillerdir. Doğal dillere örnek olarak Çince, İngilizce, Almanca, Türkçe verilebilir. Doğal Dil İşleme (DDİ) diğer adıyla Hesaplamalı Dilbilim (Computational Linguistics), bilgisayarların insan dillerinde yazılan ifadeleri veya kelimeleri anlamasını sağlamak için ayrılmış Yapay Zekâ ve dilbilimin bir alt alanıdır (Kumar 2013; Chopra ve diğ, 2013).

Dil, bir sistem, bir dizi kural ya da bir dizi sembolden meydana gelir. Semboller bir araya getirilir ve bilgi iletmek ya da bilgiyi yayınlamak amacıyla kullanılır. Kurallar ise sembollerin işlenmesinden birtakım zorluklar ortaya çıkarır. Doğal dil işleme ihtiyacı burada ortaya çıkarak bir bilgisayarın ya da makinenin, insanların yazdıklarını veya konuştuklarını (doğal dil) anlaması için ihtiyaç duyduğu her şeyi içerir (Chopra ve diğ, 2013).

Doğal dil işleme kavramını ele aldığımızda yapay zekâ ve yapay öğrenme kavramlarına da açıklık getirmemiz gerekir.

Yapay zekâ kavramı için genel bir tanım yapmak son derece güçtür çünkü zekâ kavramı için henüz net bir tanım ortaya konamamıştır. Yapay zekâ kavramı 1955 yılında Dartmouth'ta bir matematik profesörü olan ve ertesi yıl konuyla ilgili ufuk açıcı konferansı düzenleyen John

McCarthy tarafından ortaya atılmıştır (Brynjolfsson ve Mcafee, 2017). McCarthy (1989), yapay zekâ kavramını, bazı algoritma ve sözdizimsel programlamalar aracılığı ile üretilen ve insan gibi düşünme, algılama, analiz etme, deneyim kazanma gibi kabiliyetleri taklit etmeye çalışan insan elinden çıkmış sistemler olarak tanımlamaktadır. Russell ve Norvig (2021), çalışmalarında geniş anlamda yapay zekânın, bilgisayarların insan davranışını taklit etmesine ve karmaşık görevleri bağımsız olarak veya minimum insan müdahalesi ile çözmek için insan karar verme sürecini çoğaltmasına veya aşmasına olanak tanıyan herhangi bir tekniği içerdiğini ifade etmişlerdir.

Dangeti (2017), yapay öğrenme kavramını, geçmiş deneyimlerden bir şeyler öğrenmek ve gelecekteki kararları vermek için bu bilgisini kullanan bilgisayar bilimi dalı olarak tanımlamış ve yapay öğrenmenin bilgisayar bilimi, mühendislik ve istatistiğin kesiştiği noktada yer aldığını ifade etmiştir. Yapay öğrenmenin amacı, algılanabilir bir örüntüyü genelleştirmek veya verilen örneklerden bilinmeyen bir kural oluşturmaktır.

Beam ve Kohane (2018), çalışmalarında yapay öğrenmeyi yapay zekânın alt dallarından biri olarak ifade etmişler ve yapay öğrenmenin çalışma prensibinin bir durumu doğrudan programlamaktan ziyade bu durumu öğrenen ya da veriden yola çıkarak otomatik bir biçimde karar vermeyi sağlayan programlama yapısı olarak açıklamışlardır.

Bu nedenledir ki, doğal dil işleme süreci, yapay öğrenmenin çok sık kullandığı bilgisayar biliminden ayrı tutulamaz. Burada makine, insanların kullandığı doğal dilin sözdizimini ve anlamını kavrar, öğrenir ve bunun sonucunda elde edilen çıktıyı kullanıcıya sunar. Kısacası, doğal dil işleme temelde insan-bilgisayar etkileşimini basit ve etkin duruma getirmekle ilgili bir alandır (Bird ve diğ, 2009).

1.1.1 Doğal Dil İşleme Uygulama Alanları

Günlük hayatta oldukça sık karşılaştığımız doğal dil işleme teknolojileri, insan dilini anlamak için uygulamalarda gerek istatistik ve yapay öğrenme metotları gerekse kural tabanlı ve algoritmik yaklaşımlara kadar pek çok farklı yöntemden faydalanır. Genel olarak, doğal dil işleme görevleri dili temel parçalara ayırır, bu parçalar arasındaki ilişkileri kavramaya çalışır ve parçaların anlam oluştururken nasıl bir arada çalıştığını anlamaya çalışır (SAS, 2021). Bunları yaparken ise metin ve ses verilerini bazı yöntemlerle işler. Temel doğal dil işleme yöntemleri arasında tokenizasyon ve ayrıştırma, köklendirme, konuşma parçası etiketleme, dil

algılama ve anlamsal ilişkilerin belirtilmesi işlemleri bulunmaktadır. Doğal dil işleme teknolojilerine ilişkin aşağıdaki başlıklar örnek verilebilir:

Spam Filtreleme

Metin sınıflandırmasının bir uygulama alanı e-posta spam filtreleridir. İstenmeyen e-postalara karşı ilk savunma hattı olarak spam filtreleri önem kazanmaktadır. İstenmeyen posta filtrelerinin yanlış negatif ve yanlış pozitif sorunları, doğal dil işleme teknolojisinin kalbinde yer alır ve metin dizilerinden anlam çıkarma zorluğuna indirgenmiştir. Bir e-posta sistemine uygulanan bir filtreleme çözümü, gelen iletilerden hangilerinin spam olup hangilerinin olmadığını belirlemek için bir dizi protokol kullanır. İçerik filtreleri, başlık filtreleri, kural tabanlı filtreler gibi birkaç tür filtre mevcuttur. İçerik filtreleri, spam olup olmadığını belirlemek için iletinin içindeki içeriği incelemektedir. Başlık filtreleri, e-posta başlıklarında sahte bilgi aramaktadır. İzin filtreleri, mesaj gönderen herkesin alıcı tarafından önceden onaylanmasını gerektirmektedir. Genel kara liste filtreleri, kara listeye alınmış alıcılardan gelen tüm e-postaları durdurmaktadır. Kural tabanlı filtreler, belirli bir kişiden gelen postaları durdurmak veya belirli bir kelimeyi içeren postaları durdurmak gibi kullanıcı tanımlı kriterleri kullanmaktadır (Khurana ve diğ, 2017).

Makine Çevirisi

Hutchins (1986)**,** makine çevirisini "bir metni insan yardımı olmadan bir dilden (kaynak dil) başka bir dile (hedef dil) çevirmek için tasarlanmış bir bilgisayar programı" olarak tanımlamıştır. Makine çevirisi, bilgi teknolojisi unsurlarını dilbilimle birleştiren bir alan olan doğal dil işlemenin en önemli uygulamalarından birini oluşturur. İşlem sırasında veri madenciliği ve temizleme, kelime segmentasyonu, konuşma parçası etiketleme ve sözdizimsel analiz gibi birçok klasik doğal dil işleme sorununu içerir. Ayrıca makine çevirisi, yapay öğrenme algoritmalarının uygulanmasıyla da yakından ilgilidir (Jiang ve Lu, 2020).

Soru Cevaplama

Allam ve Haggag (2012), Soru Cevaplama (Question Answering-QA) sistemlerini "Bilgi Alma (Information Retrieval-IR), Bilgi Çıkarma (Information Extraction-IE) ve doğal dil işleme gibi farklı ancak ilgili alanlardan araştırmaları birleştiren bir araştırma alanı" olarak ifade etmişlerdir. Aslında, mevcut bir bilgi erişim sisteminin veya arama motorunun yapabileceği şey

sadece “belge alımı”, yani bazı anahtar kelimeler verildiğinde sadece bu anahtar kelimeleri içeren ilgili dereceli belgeleri döndürür. Bilgi erişim sistemleri yanıtları döndürmez ve buna göre kullanıcılar, yanıtları belgelerden kendileri çıkarmaya bırakılır. Ancak, bir kullanıcının gerçekten istediği şey, genellikle bir soruya kesin bir cevaptır (Hirschman ve Gaizauskas (2001); Zhang ve Lee (2003)). Yine Allam ve Haggag (2012)’a göre, soru cevaplama sistemlerinin temel amacı, şu anda çoğu bilgi erişim sisteminin yaptığı gibi, tam belgeler veya en iyi eşleşen pasajlar yerine soruların yanıtlarını almaktır.

Sohbet Robotu (Chatbot)

Chatbot, doğal dil işleme gibi tekniklerle bir metin veya ses dili ile birlikte belirli kabul edilen kurulumunda insan tartışmasını taklit eden bir Varlıktır (Dharwadkar ve Deshpande, 2018). Chatbotlar, sohbet uygulamalarında insanlarla yapılan konuşmaları simüle eden bilgisayar yazılımlarıdır. Bunun ötesinde, sadece sohbet etmekten çok daha fazlasını yapmak için büyüyebilen son derece güçlü programlardır. Chatbot'lar çok çeşitli durumlarda ve bir dizi farklı platformda kullanılır. Botlar müşteri hizmetleri, ürün önerisi, satış, planlama, pazarlama, katılım ve çok daha fazlası için kullanılır. Chatbotlar, internet üzerinden bir insan kullanıcıyla etkileşim kurmak için önceden belirlenmiş konuşma öğelerini kullanarak çalışır. Bazı sohbet robotları yapay öğrenme ve yapay zekâ kullanırken, diğerleri bu insan benzeri etkileşimleri sürdürmek için doğal dil işleme yöntemlerini kullanır (URL 1).

Özetleme

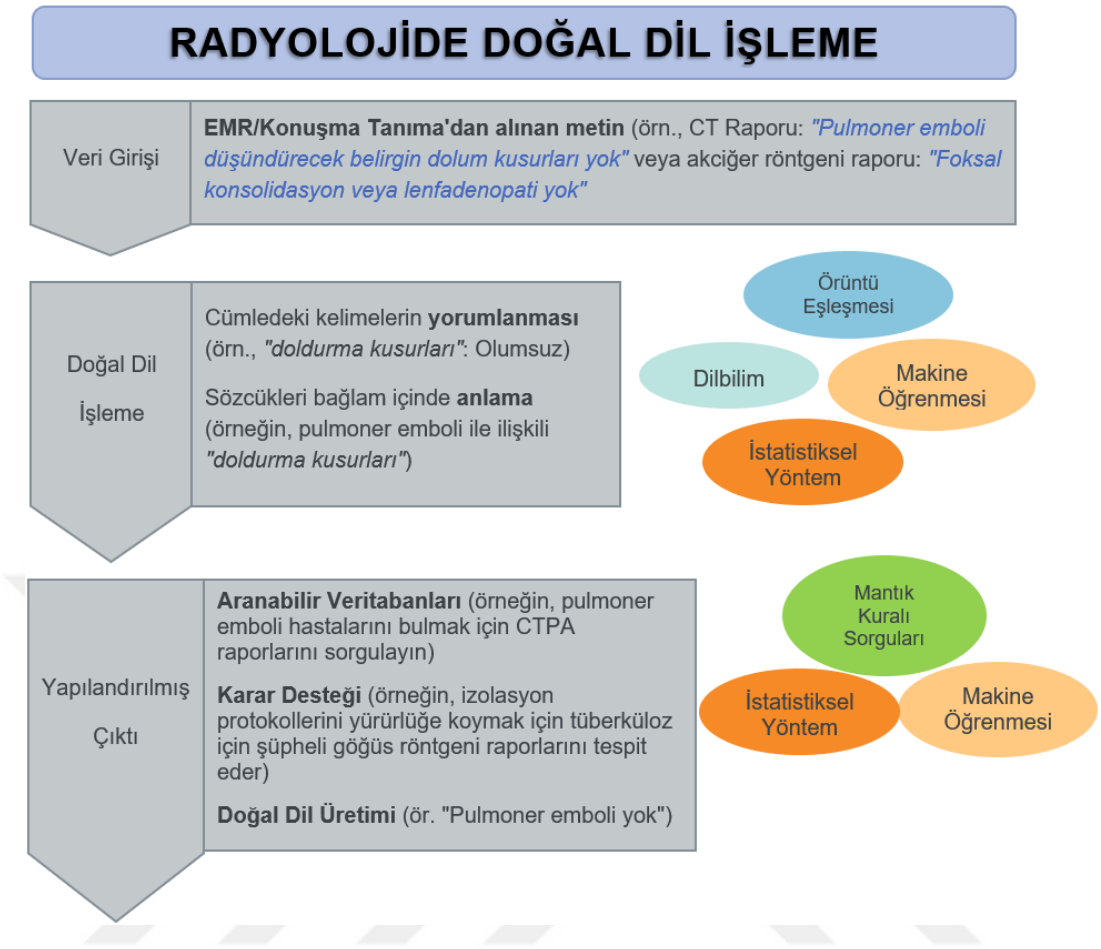
Özetleme, doğal dil işleme uygulamalarından biridir ve bilgi yoğunlaştırma için giderek daha popüler hale gelmektedir. Özetleme, bir belgenin içeriğini kullanıcının ihtiyaçlarını karşılayan yoğun bir biçimde ifade etme işlemidir. İnternette giderek daha fazla elektronik veri mevcuttur ve her şeyi okumak mümkün değildir ve bu nedenle bir tür bilgi yoğunlaştırmaya ihtiyaç vardır. Özetleme, kullanıcının muazzam miktarda bilgiden faydalı bilgileri verimli bir şekilde bulmasına yardımcı olan bir araç olarak hizmet eder (Munot ve Govilkar, 2014).

1.2 RADYOLOJİ VE DOĞAL DİL İŞLEME

Radyoloji, hastalıkları teşhis etmek ve tedavi etmek için görüntüleme teknolojisi ve radyasyon kullanan bir tıp bilimi dalıdır. Fizik, elektronik mühendisliği ve bilgisayar bilimlerindeki gelişmelerden büyük ölçüde yararlanmaktadır. Farklı tespit ve görüntüleme mantığına dayalı olarak, son yıllarda tanısal radyoloji alanında çeşitli tedavi yöntemleri ve cihazlar

geliştirilmiştir. Günümüzde hastanelerde ve tıp merkezlerinde yaygın olarak kullanılan ana tedavi yöntemi ve cihazlar arasında radyografi, floroskopi, bilgisayarlı tomografi (Computerized Tomography-CT), ultrason, manyetik rezonans görüntüleme (Magnetic Resonance Imaging-MRI) ve pozitron emisyon tomografisi (Positron Emission Tomography-PET) bulunmaktadır (Wang ve Summers, 2012).

Doğal dil işleme, zengin ve çeşitli bir klinik veri kaynağına erişim sağladığı için sağlık hizmetleri için potansiyel olarak paha biçilmez bir araçtır. Bununla birlikte, pratik klinik uygulamalar ve güncellik için kritik öneme sahiptir (Mendonça ve diğ., 2005). Doğal dil işleme, radyoloji raporlarını analiz etmek için giderek daha fazla kullanılmaktadır (Cai ve diğ. (2016); Yim ve diğ., (2016)). Radyoloji raporları, bireysel hasta bakımı (Sahni ve Khorasani, 2016) ve sağlık sistemlerinin kalite iyileştirmesi için değerlidir (Jay ve diğ., 2017). Toplu düzeyde, anonimleştirilmiş radyoloji raporları tanı verimini değerlendirmek, kılavuza uyumu değerlendirmek, epidemiyolojik araştırma yapmak ve akran geri bildirimi ve klinisyen sevk geribildirimi için kullanılabilir (Goel ve diğ, (2019); Pons ve diğ, (2019)). Ancak bu uygulamalar yaygın olarak uygulanmamaktadır, çünkü serbest metin radyoloji raporlarının manuel olarak sınıflandırılması hantaldır (Spasic ve diğ, 2020). Radyolojide, doğal dil işleme, radyoloji raporlarından değerli bilgilerin çıkarılmasını sağlar. Kalite iyileştirme, epidemiyolojik araştırma ve kılavuza uyumu izleme gibi çeşitli alt görevler için kullanılabilir (Olthof ve diğ, 2021).



Şekil 1.1 Radyolojide Doğal Dil İşleme (Türkçeye uyarlanmıştır) (Cai ve diğ, 2016).

Şekil 1.1'de özetlendiği şekilde günümüz radyolojisinde anlaşıldığı şekliyle doğal dil işlemenin, doğal dilden bilgi çıkarmayı amaçlayan çeşitli tekniklerin bir koleksiyonu olduğu gösterilmektedir (örneğin, ilgilenilen kavramları çıkarmak ve bunları yapılandırılmış bir formata koymak için bir radyoloji raporunu analiz etmek). Ayrıca bu çıktıyı raporları aranabilir bir veritabanında indekslemek, hasta veya rapor düzeyinde sınıflandırma sağlamak veya bulguları daha basit bir doğal dilde özetlemek için kullanılır (Cai ve diğ., 2016).

2.GENEL KISIMLAR

Bu bölümde bibliyografik çalışma ve bilgi çıkarımı kavramı, bilgi çıkarımını oluşturan analizler, konu modelleme ve konu modelleme kapsamında kullanılan yöntemler ayrıntılı bir şekilde incelenmektedir. Ayrıca ilgili konularda gerçekleştirilen literatür taraması da yine bu bölümün sonunda yer almaktadır.

2.1 BİBLİYOGRAFİK ÇALIŞMA

Bibliyografi kelimesi Yunanca “kitap” anlamına gelen “bibliyo” ve “yazmak” anlamına gelen “grafein” kelimelerinin birleştirilmesinden meydana gelmiştir. Bibliyografiler bize basılı eserler hakkında bilgi vermektedir. Bibliyografik çalışmaların amacı bir konu hakkında çalışma yapılacağı zaman tekrara düşmemek ve özgün bir eser ortaya koymaktır (Kılıçarslan ve Altuğ, 2016). Bu tez çalışmasında da özellikle son yıllarda sağlık alanında makine öğrenmesi çalışmalarının hangi çalışma alanlarında yoğunlaştığını hangi konulara yöneldiği araştırılarak bir bibliyografik çalışma yapılmak istenmiştir.

2.2 BİLGİ ÇIKARIMI

Grishman (2015), bilgi çıkarımının (IE – Information Extraction) amacını, metnin anlamsal yapısını, onu kullanabilmemiz için açık hale getirmek olduğunu ifade etmiştir. Belirtilen bu ifade metni analiz etme ve içindeki anlamsal olarak tanımlanmış varlıkların ve ilişkilerin sözlerini belirleme sürecidir. Bu ilişkiler daha sonra belirli ilişkileri aramak veya açıkça belirtilen gerçeklerden ek bilgiler çıkarmak için bir veri tabanına kaydedilebilir.

Bilgi çıkarımı, çalışılacak metinsel veri hacmi ve anlamsal ilgi ilişkilerinin yüksek olduğu durumlarda oldukça önemli bir uygulamadır. Yine Grishman (2015)’ın ifadesine göre, tıbbi ve biyomedikal literatür, yılda 500.000'den fazla makale hızında büyümekte ve hastaneler ve tıbbi uygulamalar, her hasta ziyaretinde veya kabulünde gözden geçirilecek büyük hacimli elektronik tıbbi kayıtlar üretmektedir. Bu gibi büyük hacimli verilerin bilgi çıkarımı ile özetlenmesi geleneksel araştırma yöntemlerine göre araştırmacıya zaman kazandırmakta ve daha kısa sürede faydalı bilgilere ulaşmasında önemli rol oynamaktadır.

Bilgi çıkarımı biyomedikal araştırmalardan finansa kadar oldukça geniş bir kullanım alanına sahiptir. Jiang (2012), bu kullanım alanlarına örnek olarak şunlar vermiştir:

- Biyomedikal araştırmalarda belirli genler, proteinler veya diğer biyomedikal varlıklarla ilgili keşifleri aramak anahtar kelime eşlemeye dayalı basit arama yeterli olmayabilir. İlgili literatüre özel terimler alana özgü mevcut bilgi tabanlarındaki karşılık gelen girişler ile eşleştirilerek bilgi çıkarımı sağlanır**.**
- Finans alanında ise uzmanlar günlük karar vermelerine yardımcı olmak için genellikle haber makalelerinden belirli bilgi parçaları aramaya ihtiyaç duyarlar. Metinden bu tür bilgileri otomatik olarak bulmak, adlandırılmış varlık tanıma ve ilişki çıkarma gibi standart bilgi çıkarma teknolojilerini gerektirir.
- İstihbarat analistleri, terör olaylarını tanımlayan belgeleri hızlı bir şekilde bulmak için bilgi alma teknolojilerini kullanılabilirken, bu belgelerdeki belirli bilgi birimlerini daha fazla saptamak için bilgi çıkarma teknolojilerine ihtiyaç duyarlar.
- Bunların yanı sıra her gün bilgi almak için kullandığımız arama motorları tek kelimelik aramalar yerine varlık arama, yapılandırılmış arama ve soru yanıtlama gibi daha gelişmiş arama sorunları, kullanıcılara daha iyi arama deneyimi sağlayabilir. Bu arama yeteneklerini kolaylaştırmak için, genellikle belge temsilini zenginleştirmek veya temel alınan bir veritabanını doldurmak için bir ön işleme adımı olarak bilgi ayıklamaya ihtiyaç duyulur.”

Bilgi çıkarımı birkaç analizin bir araya gelmesinden oluşmaktadır. Bu analizlere Adlandırılmış Varlık Tanıma (Named Entity Recognition (NER)), Sözdizimsel Analiz (Syntactic Analysis), Eşreferans Çözünürlüğü (Co-referance Resolution (CO)), Anlamsal Analiz (Semantic Analysis), İlişki Çıkarımı (Relation Extraction (RE)) ve Olay Çıkarımı (Event Extraction (EE)) örnek gösterilebilir. Bilgi çıkarımını oluşturan bu analizlerin uygulama olarak aynı sırayla yapılması gerekmez.

Bu analizlerin nerelerde ve hangi amaçlarla kullanıldığı ayrıntılı bir şekilde başlıklar altında ele alınmıştır:

2.2.1 Adlandırılmış Varlık Tanıma (Named Entity Recognition (NER))

Adlandırılmış Varlık Tanıma (NER); kavramı kuruluşlar (Dünya Sağlık Örgütü), kişiler (Mustafa Kemal Atatürk), yer adları (Hint Okyanusu), zamansal ifadeler (1 Eylül 2011), sayısal ve para birimi ifadeleri (50 Milyar Dolar) gibi küçük ölçekli bir şablonun doldurulması yoluyla belirlenen varlıkları ifade etmektedir. Örneğin, kişiler söz konusu olduğunda, kişinin unvanı,

konumu, uyruğu, cinsiyeti ve diğer özelliklerinin çıkarılmasını içerebilir (Marrero ve diğ, 2013).

2.2.2 Sözdizimsel Analiz (Syntactic Analysis)

Brill ve Mooney (1997) çalışmalarında sözdizimsel analizi, bir cümlenin dilbilgisel yapısının, yani kelimelerin isim cümleleri ve fiil cümleleri gibi bileşenlere nasıl gruplandırıldığının belirlenmesi olarak tanımlamışlardır.

2.2.3 Anlamsal Analiz (Semantic Analysis)

Anlamsal analiz, doğal dil işleme yaklaşımının temel bir özelliğidir. Bir cümlenin veya paragrafın bağlamını uygun biçimde belirtir. Semantik, dilin önemi çalışması ile ilgilidir. Kullanılan kelime dağarcığı, dil sınıfları arasındaki karşılıklı ilişkiden dolayı konunun önemini aktarmaktadır (Velupillai ve diğ, 2015).

2.2.4 Eş referans Çözünürlüğü (Co-reference Resolution (CO))

Kong vd (2010), Eş referans Çözümlemesi kavramını, hangi isim tamlamalarının (zamirler, özel isimler ve ortak isimler dahil) belgelerdeki aynı varlıklara atıfta bulunduğunu belirleyen bir uygulama olarak tanımlamışlardır. Örneğin, cümlede, “Ben geldim ama burada kimse yok” dedi. Bu cümlede “Ben” kişinin adını ve “Burada” bulunduğu yeri ifade etmektedir. Bu nedenle, referans çözünürlüğü, doğal dil anlama, özetleme, bilgi çıkarma, metinsel gereklilik ve benzeri görevlerde hayati bir rol oynar (Singh, 2018).

2.2.5 İlişki Çıkarımı (Relation Extraction (RE))

İlişki çıkarımı, metin kaynaklarından bir veritabanı şemasıyla eşleşen üçlüler (özne varlığı, yüklem ilişkisi, nesne varlığı) biçimindeki yapılandırılmış verileri kurtarmayı amaçlayan doğal dil işleme alanıdır. Örneğin, "Microsoft, Internet Explorer tarayıcısını etkileyen kritik bir hata için bir düzeltme yayınladı." cümlesinden, (Microsoft, Internet Explorer'ın satıcısıdır) örneğini (yazılım satıcısı, satıcı, yazılım ürünü) ilişkisi ortaya çıkarılır (Jones, 2015).

2.2.6 Olay Çıkarma (Event Extraction (EE))

Olay çıkarmanın amacı, metinlerdeki olay örneklerini tespit etmek ve varsa, olay türünü, tüm katılımcılarını ve niteliklerini tanımlamaktır. Farklı olay türleri farklı argümanlarla tanımlanabilse de, olay çıkarmanın basit bir özeti, yapılandırılmamış doğal dillerden, olayların yapılandırılmış temsilini elde etmektir. Haber makaleleri, sosyal medya gönderileri vb. gibi çok

sayıda metin kaynağından gerçek dünya olaylarının "kim, ne zaman, nerede, ne, neden, nasıl" dahil olmak üzere "5W1H" (Who, What, When, Where, Why, How) sorularını yanıtlamaya yardımcı olmak için kullanılır (Xiang ve Wang, 2019).

Bilgi çıkarımı ile ilgili bir diğer konu olan Konu Modelleme ayrı bir başlık altında ele alınmıştır.

2.3 KONU MODELLEME

Gizli (yani, doğrudan gözlemlenemeyen) konuları çıkarma ve öğrenme veya bir metin bütünü boyunca gizli tematik bilgileri ortaya çıkarma süreci, *konu modelleme (topic modeling)* olarak adlandırılır. Konu modellemenin temelindeki varsayımlar şunlardır: her konu bir kelimeler topluluğudur ve her belge konuların bir karışımı olarak temsil edilebilir (Vangara, 2020). Kısaca konu modeli, bir belge koleksiyonundan gizli anlambilimi çıkarabilen istatistiksel bir modeldir (Deerwester ve diğ, 1990). Konu modelleme teknikleri, doğal dil işlemeden anlamsal madenciliğe ve belgelerde ve veri kümelerinde gizli keşifte çok yararlı ve etkili olabilir (Jelodar ve diğ, 2019). Konu modelleme yöntemleri genellikle büyük elektronik arşivleri otomatik olarak düzenlemek, anlamak, aramak ve özetlemek için kullanılır. "Konular", bir kelime dağarcığındaki sözcükleri ve bunların belgelerdeki oluşumlarını birbirine bağlayan gizli, tahmin edilmesi gereken değişken ilişkileri ifade eder. Bir belge, konuların bir karışımı olarak görülür. Tong ve Zhang (2016), konu modellerini, koleksiyon boyunca gizli temaları keşfeder ve belgelere bu temalara göre açıklamalar ekler şeklinde açıklamışlardır. Her kelime bu konulardan birinden alınmış olarak görülür. Son olarak, konuların bir belge kapsamı dağılımı oluşturulur ve konuların perspektifiyle ilgili verileri keşfetmek için yeni bir yol sağlar.

Kısaca söylemek gerekirse, olasılıksal konu modelleri bir özeti özetler. Bir konu modeli geliştirmenin önemli bir yönü, farklı kelimeler arasındaki kullanım ve anlamlarının ötesine geçen bir bağlantı olarak tanımlanan kavramlar arasındaki benzerlik derecesini belirlemektir. Önceki yayınlar, örneğin FDA önceliklerini keşfetmek veya tıbbi notlardaki kavram ile genetik bilgi arasındaki ilişkiyi değerlendirmek için bir bibliyografik analiz gibi çeşitli senaryolarda kullanımını araştırmıştır. Bir konu, işlevsel olarak, bir kelime grubunun bir dizi kelime öbeği üzerinde bir araya gelme olasılığı olarak tanımlanır. Konu-dosya dağılımları, kelime-konu dağılımları ve gizli parametreler daha sonra model tarafından gözlemlenen kelimeler ve belgeler aracılığıyla tahmin edilir. Latent Dirichlet Allocation (LDA) modeli, doğal dil çalışmasında kullanılan, bir dizi kaynak belgeden konuların çıkarılmasına ve belgelerin tek tek bölümlerinin benzerliği hakkında mantıklı bir açıklama sağlanmasına izin veren üretken bir

modeldir. LDA'nın üretken süreci, metinde bulunan verilerin analizine (metin madenciliği) dayanır. Kelime kombinasyonları rastgele değişkenler olarak kabul edilir (Sperandeo ve diğ, 2020).

Konu modellemenin uygulama alanları dört başlık altında incelenebilir:

2.3.1 Gizli Anlamsal Analiz

Gizli Anlamsal Analiz (Latent Semantic Analysis-LSA), kelimelerdeki ve pasajlardaki anlamı, temeldeki kavramların doğrusal kombinasyonları olarak tahmin etmek için istatistiksel bir yöntemdir. Bu temel kavramlar, gözlemlenen sözcük kullanım kalıplarının matris işlemleri yoluyla çıkarılır (Kulkarni ve diğ, 2014).

LSA, belgelerde, paragraflarda veya cümlelerde kullanım bağlamları aracılığıyla kelimeler arasındaki ilişkileri çıkarmaya yönelik tam otomatik bir istatistiksel yaklaşımdır. Morfolojik, sözdizimsel veya anlamsal ilişkileri analiz etmek için doğal dil işleme tekniklerini kullanmaz. Sözlükler, eş anlamlılar sözlüğü, sözlüksel referans sistemleri, anlamsal ağlar veya diğer bilgi temsilleri gibi insan tarafından oluşturulmuş kaynakları da kullanmaz. Tek girdisi büyük miktarda metindir. LSA bir "denetimsiz öğrenme" tekniğidir. Geniş bir metin koleksiyonu ile başlar, bir terim-belge matrisi oluşturur ve bilgi alımı ve ilgili metin analizi problemi için yararlı olan bazı benzerlik yapılarını ortaya çıkarmaya çalışır (Dumais, 2004).

2.3.2 Negatif Olmayan Matris Çarpanlarına Ayırma

Negatif olmayan matris çarpanlara ayırma (NonNegative-Matrix Factorization-NMF), bir boyut küçültme yöntemi ve faktör analizi yöntemidir. Birçok boyut indirgeme tekniği, matrislerin düşük dereceli yaklaşımlarıyla yakından ilişkilidir ve NMF, düşük dereceli faktör matrislerinin yalnızca negatif olmayan elemanlara sahip olacak şekilde sınırlandırılması bakımından özeldir. Belge verileri için bir kümeleme ve konu modelleme yöntemi olarak kullanılır (Choo ve diğ, 2013). Gizli anlamsal analiz ve negatif olmayan matris çarpanlarına ayırma, metin tümcesinin sözcük torbası (Bag of Words-BoW) veya TF-IDF'sini (Term Frequency-Inverse Document Frequency) kullanarak tümlemin matris tabanlı bir temsilini ayrıştıran faktör analizini takip eder (Vangara ve diğ, 2020).

2.3.3 Gizli Dirichlet Ayrımı

Gizli Dirichlet Ayrımı (Latent Dirichlet Allocation-LDA), belgelere konular atar ve bir metin koleksiyonu verilen kelimeler üzerinde konu dağılımları oluşturur. Bunu yaparken, kelimeler arasındaki benzerlikle ilgili herhangi bir yan bilgiyi yok sayar. Bununla birlikte, konular arasında şaşırtıcı derecede yüksek bir tutarlılık kalitesine ulaşır (Blei ve diğ, 2003). Bir derlemi modellemek için denetimsiz bir üretken olasılık yöntemi olan LDA, en yaygın olarak kullanılan konu modelleme yöntemidir. LDA, her belgenin gizli konular üzerinde olasılıksal bir dağılım olarak temsil edilebileceğini ve tüm belgelerdeki konu dağılımının ortak bir Dirichlet'i paylaştığını varsayar. LDA modelindeki her bir gizli konu aynı zamanda kelimeler üzerinde olasılıksal bir dağılım olarak temsil edilir ve konuların kelime dağılımları da ortak bir Dirichlet'i paylaşır (Jelodar ve diğ, 2019). LDA, her belgeyi aynı anda birden çok konuyu ifade edecek şekilde modeller; bir belgenin her konuyu belirli bir yakınlıkla ifade ettiği söylenir. Aynı şekilde, her konu kelimeler üzerine bir dağılımdır. Bu nedenle, matematiksel açıdan her belge bir dağılım karışımıdır. Söz konusu ve konu belge matrislerini (konularda görünen kelimelerin ve belgelerde görünen konuların olasılıkları) bulmak için, bu dağılımlarla ilk belge kümesine yaklaşmak gerekir (Koltcov ve diğ, 2014). En popüler iki yaklaşım, sırasıyla değişken yaklaşımlara (Blei, 2011; Blei ve diğ, 2003) ve Gibbs örneklemesine (Griffiths ve Steyvers, 2004) dayanmaktadır. Bu algoritmalar, veri kümesinin ortak olabilirlik fonksiyonunun yerel bir maksimumunu bulur; bu konu modelleme problemi için bir çözüm olarak kabul edilmektedir (Koltcov ve diğ, 2014).

LDA, $D=\{d_m\}$, $m \in [1, ..., M]$ olmak üzere bir belge koleksiyonu oluşturur.

k konusunun söz varlığına dağılımı $\Phi=\{\phi_k\}$, $k \in (1, \ldots, K)$ olarak ve m-inci belgenin tüm K konularına dağılımı $\Theta=\{\theta_m\}$, $m \in [1, ..., M]$ olarak gösterilir.

Konu, terimlerin bir kelime dağarcığı üzerindeki dağılımıdır. Her belgenin, şu şekilde ifade edilebilecek konulara göre dağılım olarak tanımlanmasına olanak tanır (Guan ve diğ, 2019):

$$P(w|d) = P(w|t) \times P(t|d) \quad (2.1)$$

Burada *w* bir kelimeyi temsil eder, d bir belgeyi temsil eder ve *t* konuyu temsil eder. Aşağıdaki gibi genişletilebilir:

$$p(w, z, \theta_m, \Phi|\alpha, \beta) = p(\Phi|\beta)p(\theta_m|\alpha)p(z|\theta_m)p(w|\Phi, z) \quad (2.2)$$

Burada m belgesi için belgelerin θ_m konuları üzerinden dağılımı ve konuların kelime dağarcığı Φ üzerinden dağılımı sırasıyla α ve β önceliklerinden örneklenir. Daha sonra, her bir kelime için *z* konu ataması θ_m'den üretilir ve doğru *w* sözcükleri ilgili konu atamalarına *z* ve konuların Φ kelime dağarcığına göre dağılımına göre oluşturulur (Guan ve diğ, 2019).

2.3.4 Pachinko Dağılım Modeli

Li ve McCallum (2006), Pachinko Dağılım Modeli (Pachinko Allocation Model-PAM)'ni açıklarken, tek bir konu grubunun dağılımını bir grafikte belgeler ve birlikte oluşumları temsil ettiğini ifade etmişlerdir. PAM'de her düğüm, bir sonraki alt seviyedeki düğümler üzerindeki bir dağılımı temsil eder.

İlgili yöntemlere ait benzer çalışmaları içeren literatür taraması ise şu şekildedir:

Amado ve diğ. (2017), çalışmalarında 2010-2015 döneminde gerçekleştirilen Pazarlamada Büyük Veri üzerine bir araştırma literatürü analizini özetlemektedir. Bu zaman dilimini, bu yıllarda Web'in Büyük Veri'ye olan ilgisinin artmasına göre seçildiğini ifade etmişlerdir. Özellikle, analiz beş boyutla ilgili terimlere ve konulara odaklanmaktadır: Büyük Veri, Pazarlama, Yazarların bağlantılarının coğrafi konumu (ülkeler ve kıtalar), Ürünler ve Sektörler. Çalışmalarında kullandıkları makalelerde analiz edilen bu boyutlarla ilgili sözlükler oluşturulmuştur. Bu sözlükler, aynı anlama gelen farklı kelimelerin tek bir kelime ile ifade edilmesi için önemlidir (örneğin "big data" ve "massive data"). Metin madenciliğinde kullanılan stemming (köklendirme) işlemi uygulanmıştır. Çalışmada metin madenciliğinde hesaplamalı deneyler için R programında "tm" (text mining) paketi, Gizli Dirichlet Ayrımı için de "topicmodels" paketi kullanılmıştır. LDA belge terim matrisi ile beslenmiştir.

Chen ve diğ. (2021), çalışmalarında son 25 yılda yayınlanan (2020 yılı son yıl) 1295 makaleyi inceleyerek bilgisayar destekli dil öğrenimi (computer-assisted language learning-CALL) konusunu ele almışlardır. CALL'daki araştırma durumunu, eğilimleri ve öne çıkan sorunları araştırmak için yapısal konu modelleme, Mann-Kendall eğilim testi ve bibliyometri ile hiyerarşik kümeleme yöntemlerini kullanmışlardır.

Abdelaziz ve diğ. (2021), çalışmalarında iki amaca yönelik sistematik bir literatür incelemesi sunmuşlardır: binaların enerji tüketimini etkileyen ilgili faktörleri anlamak ve farklı bina türlerinin enerji tüketimini sınıflandırabilen ve tahmin edebilen en iyi akıllı hesaplama yöntemlerini bulmak. Araştırma için 2013'ten 2020'ye kadar 822 makale incelenmiş ve başlık,

özet taraması ve deneyli makalelere dayalı olarak 106 makaleye odaklanmışlardır. Makale, PRISMA metodolojisini, metin madenciliği yaklaşımını ve bibliyometrik harita analizini (Vosviewer ile) kapsamaktadır. 106 makale manuel analiz için çok sayıda olduğu için metin madenciliğinden yararlanılmıştır. Bu sebeple, metin madenciliği yöntemlerinin uygulanabilmesi için yine bu çalışmada da araştırma konusuyla ilgili kavramları içeren bir sözlükten yararlanılmıştır.

Kane ve diğ. (2016), çalışmalarında buğday, pirinç, çavdar, sorgum ve güvercin bezelyesi dahil olmak üzere çok yıllık temel mahsuller üzerine yapılan araştırmaların gelişimini karşılaştırmışlardır. Bunun için Web of Science, Scopus, ScienceDirect ve Agricola'yı kullanarak 1930'dan 2016 yılına kadar yayınlanmış 914 makaleyi incelemişlerdir. Burada araştırma ve yayıncılıktaki eğilimleri anlamak için tüm kütüphanedeki ve her ürünle ilgili literatür koleksiyonlarındaki meta verileri analiz etmişler ve makale özetlerine birlikte ortaya çıkan terimleri ve gizli konuları tanımlayan bir tür metin madenciliği yöntemlerinden biri olan konu modellemesi uygulamışlardır.

Jiang ve diğ. (2015), çalışmalarında hidroelektrikle ilgili küresel bilimsel literatürü 1994'ten 2013'e kadar nicel olarak değerlendirmek için konu modelleme tabanlı bibliyometrik analiz kullanmışlardır. 1726 bilimsel makaleyi analiz etmişlerdir. Bu çalışmada, modelleme süreci Topicmodels R paketine dayanmaktadır. Ayrıca istatistiksel verilerin işlenmesi ve grafiklerin oluşturulması için MS Excel'den yararlanılmıştır.

Porturas ve Taylor (2020), çalışmalarında son 40 yılda acil tıpta en yaygın araştırma temalarını konu modellemeye yönelik denetimsiz, makine öğrenimi yaklaşımıyla belirlemeyi ve eğilimlerini ve özelliklerini özetlemeyi amaçlamışlardır. Bunun için 1980'den 2019'a kadar acil tıp alanındaki tüm araştırma makaleleri, makale özetleri de dahil olmak üzere eksiksiz referans girişlerini, yaygın olarak alıntılanan altı dergi için derlemişlerdir. Özetleri doğal dil işleme yöntemleri ve denetimsiz konu keşfi için ise Gizli Dirichlet Ayrımı kullanılarak analiz etmişlerdir.

Audrin ve Audrin (2022), çalışmalarında dijital okuryazarlığın öğrenme ve eğitim alanındaki araştırma alanına genel bir bakış sağlamayı amaçlamışlardır. Metin madenciliğini kullanarak, konuyla ilgili 2000 ile 2020 yılları arasında yayınlanmış 1037 araştırma makalesini incelemişlerdir. Metin madenciliği kullanarak analizi gerçekleştirmek için WordStat 8 yazılımı kullanılmıştır. Analizde hem köklendirme hem de lemmatizasyon işlemleri gerçekleştirilmiştir

Kelimelerin frekanslarının doğru belirlenmesi için bir sözlük kullanılmıştır. Ayrıca analiz sürecinde faktör analizi ve çok boyutlu ölçeklemeden yararlanılmıştır.

Karami ve diğ. (2020), çalışmalarında amaçları Twitter tabanlı araştırmaların baskın konularını belirlemek, konuların zamansal eğilimini özetlemek ve konuların son on yıldaki gelişimini yorumlamaktır. Bunun için Mart 2019'da Web of Science (WOS), EBSCO ve IEEE olmak üzere üç ana veri tabanından başlığında "twitter" geçen ilgili özetlere erişilmiştir. 2006 ve 2019 yılları arasında yayınlanan dergi ve konferans özetleri ele alınmıştır. İlk olarak genel bir bakış açısına sahip olmak için, sırasıyla çubuk grafiği ve kelime bulutunu kullanarak ilk 10 ve ilk 50 kelimenin sıklığı analiz edilmiştir. Konu modelleme için Gizli Dirichlet Ayrımı kullanılmıştır.

Zou (2018), çalışmasında ilaç güvenliği konusundaki araştırma eğilimlerini analiz etmek için 2007'den 2016'ya kadar uyuşturucu güvenliğine adanmış dört dergideki 4347 makalenin başlıklarını ve özetlerini incelemiştir. 50 ana konuyu çıkarmak için Gizli Dirichlet Ayrımı modelini kullanmış ve bunların zamansal popülerliğini keşfetmek için trend analizi yapmıştır.

3. MALZEME VE YÖNTEM

3.1 VERİ SETİ

PubMed bibliyografik veri tabanı kullanılarak makine öğrenimi (machine learning) ve radyoloji (radiology) ile ilgili yayınlardan başlıkları ve özetleri araştırılarak veri seti oluşturulmuştur. Tez çalışmasının veri tabanı makine öğrenmesi kavramının radyoloji alanında kullanımı ile ilgili çalışmalarla sınırlandırmak amacıyla PubMed tıbbi veri tabanı üzerinden yapılmıştır. PubMed biyomedikal bir veri tabanı olup içerisinde başvuru kitapları, moleküler biyoloji, genetik ve tıp bilimleri ile ilgili konulara ulaşılabilir. Sağlık bilimleri alanında uluslararası makaleler ve son gelişmeler bu veri tabanı aracılığı ile takip edilebilir.

Araştırmamız şu yaklaşımı içermektedir: makine öğrenimi VE radyoloji [Başlık/Özet] VE ("2017/01/01"[Date-Entrez]:"2023/03/31"[Date-Entrez]) VE özet[metin]. Şekil 3.1'de PubMed veri tabanından veri çekmede kullanılan alan kısaltmaları verilmiştir (https://www.nlm.nih.gov/bsd/mms/medlineelements.html). Bu tez çalışmasında PubMed Unique Identifier (PMID), Abstract (AB), Date Publication (DP) ve Title (TI) alanları kullanılmıştır.

Field	Abbreviation
Abstract	(AB)
Copyright Information	(CI)
Affiliation	(AD)
Investigator Affiliation	(IRAD)
Article Identifier	(AID)
Author	(AU)
Author Identifier	(AUID)
Full Author	(FAU)
Book Title	(BTI)
Collection Title	(CTI)
Comments/Corrections	
Conflict of Interest Statement	(COIS)
Corporate Author	(CN)
Create Date	(CRDT)
Date Completed	(DCOM)
Date Created	(DA)
Date Last Revised	(LR)
Date of Electronic Publication	(DEP)
Date of Publication	(DP)
Edition	(EN)
Editor and Full Editor Name	(ED) (FED)
Entrez Date	(EDAT)

Field	Abbreviation
Gene Symbol	(GS)
General Note	(GN)
Grant Number	(GR)
Investigator Name and Full Investigator Name	(IR) (FIR)
ISBN	(ISBN)
ISSN	(IS)
Issue	(IP)
Journal Title Abbreviation	(TA)
Journal Title	(JT)
Language	(LA)
Location Identifier	(LID)
Manuscript Identifier	(MID)
MeSH Date	(MHDA)
MeSH Terms	(MH)
NLM Unique ID	(JID)
Number of References	(RF)
Other Abstract	(OAB)
Other Copyright Information	(OCI)
Other ID	(OID)
Other Term	(OT)
Other Term Owner	(OTO)
Owner	(OWN)

Field	Abbreviation
Pagination	(PG)
Personal Name as Subject	(PS)
Full Personal Name as Subject	(FPS)
Place of Publication	(PL)
Publication History Status	(PHST)
Publication Status	(PST)
Publication Type	(PT)
Publishing Model	(PUBM)
PubMed Central Identifier	(PMC)
PubMed Central Release	(PMCR)
PubMed Unique Identifier	(PMID)
Registry Number/EC Number	(RN)
Substance Name	(NM)
Secondary Source ID	(SI)
Source	(SO)
Space Flight Mission	(SFM)
Status	(STAT)
Subset	(SB)
Title	(TI)
Transliterated Title	(TT)
Volume	(VI)
Volume Title	(VTI)

Şekil 3.1 Veri Çekmede Kullanılan Alan Kısaltmaları

Metinsel içerik daha sonra sık kullanılan kelimelerin (örneğin, arka plan, amaç, yöntem, sonuç, sonuç, bilgi, sık, bağlam, arasında, önermek, birlikte, mümkün, içermek, makale, ayrıca, göstermek, daha sonra, ana, görüntülemek, içinde ve bulmak) sayısal rakamlar, durdurma sözcükleri ve noktalama işaretlerinin elenmesiyle ön işleme tabi tutulmuştur. Ön işleme Python programlama dili aracılığıyla gerçekleştirilmiştir.

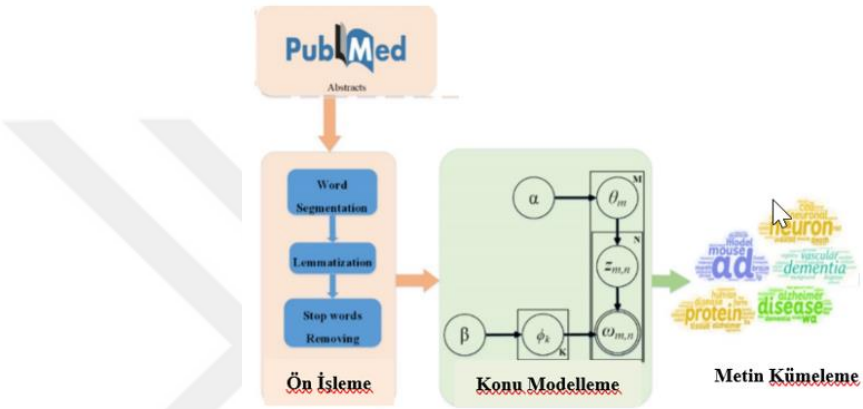
3.2 VERİ ÖNİŞLEME

Özet, bir makalenin önemli içeriğinin özlü ve doğru bir tanımını sağlayabileceğinden, tam metin okunmadan özetlerden gerekli bilgiler elde edilebilmektedir. PubMed'de makale aramak için giriş terimi olarak "makine öğrenimi ve radyoloji" seçilerek ve 2017 ile 2023 yılları arasında 11385 makale elde edilmiştir.

Her dosya yarı yapılandırılmış bir XML belgesidir ve <title>, <abstracts>, <pmid> vb. gibi çeşitli etiketler içerir. <abstract> ve <pmid> alanlarındaki içeriği ham XML dosyalarından çıkarılmıştır. PMID (PubMed Unique Identifier) yani, PubMed arama motorunda her çalışma

için tekil bir tanımlayıcıdır. Ayıklamadan sonra, her özet karşılık gelen bir dosyada saklanmıştır.

Tez çalışmasında öncelikle PubMed veri tabanında anahtar kelimelere uygun; özetleri mevcut makaleler taranmıştır. Bu makaleler içerisinden 2017-2023 yılları arasında yayımlanmış olanların özet, makale başlığı, yazar bilgileri ve yılları ile birlikte indirilerek veri seti oluşturulmuştur. Bunun için Python kütüphanelerinden yararlanılmıştır. İlgili kütüphaneler bölüm 3.3'te açıklanmaktadır. Ardından Şekil 3.2'deki adımlar izlenmiştir.



Şekil 3.2: Çalışma Metodolojisi (Guan ve diğ, 2019)

Şekil 3.2'de de görüldüğü gibi metin verilerinin işlenmesi için ilk olarak önişleme adımının gerçekleştirilmesi gerekmektedir. Bu adımlardan bazıları şu şekildedir:

Durdurma sözcüklerini kaldırma: Durdurma sözcükleri doğal dillerin bir kısmını oluşturmaktadır. Bu sözcüklerin metinden çıkarılmasının nedeni, terim uzayının boyutunun azaltılmasıdır. Metin verilerinde en sık kullanılan kelimeler, belgelerin anlamını vermeyen edatlar, zamirler vb. durdurma sözcüklerini oluşturmaktadır. Durdurma sözcükleri için örnek olarak peki, belki, sence, ama, aynen kelimeleri verilebilir. Durdurma sözcükleri, metin madenciliği uygulamalarında anahtar sözcükler olarak ölçülmediği için belgeden kaldırılır (Porter, 1980).

Tokenizasyon: Tokenizasyon, bir metin akışını sözcüklere, ifadelere, sembollere veya belirteç adı verilen diğer anlamlı öğelere ayırma işlemidir. Tokenizasyonun amacı, bir cümledeki sözcüklerin keşfedilmesidir (Kannan ve diğ, 2014).

Köklendirme: Köklendirme, bir kelimenin değişken biçimlerini ortak bir temsil olan kökte birleştirme işlemidir (Kannan ve diğ, 2014). Köklendirme işleminde kelime köklerinin elde edilmesi için kelimedeki son ekler çıkarılır ve böylece önemli bir bilgi kaybı yaşanmadan karmaşıklığı azaltır (Doğuç, 2022).

3.3 VERİ ANALİZ

3.3.1 Veri Temizleme ve Ön İşleme

Veri temizleme ve ön işleme aşamasına geçmeden önce ilk olarak gerekli kütüphaneler indirilmiştir. Bu kütüphanelerden "pandas" veri analiz ve manipülasyonu için, "re" düzenli ifadelerle çalışmak için, "matplotlib" ve "seaborn" görselleştirme için, "nltk" doğal dil ile analiz yapmak için, "numpy" matematiksel işlemler için, "gensim" denetimsiz konu modelleme ve doğal dil işleme analizleri için, "word cloud" ise yine görselleştirme için kullanılmaktadır.

Daha sonra çekilen özetlerdeki eksik veriler ve yinelenen değerler kaldırıldı. İngilizcede yer alan kısaltmaların ("he's : he is", "can't : cannot" gibi) yer aldığı bir sözlük oluşturularak bu kelimelerden arındırılmıştır.

Bu işlemlerin ardından özetlerdeki kelimelerde bulunan tüm harflerin analiz aşamasında sorun yaratmaması adına küçük harfe dönüştürülmesi işlemi gerçekleştirilmiştir.

Harf dönüşümlerinden sonra veriden "Semboller" tablosunda yer alan karakterlerin silinmesi işlemi yapılmıştır. Bu işlem yapılırken "re" modülünden yararlanılmıştır. "re" yani "Regular Expression" metindeki düzenli ifadelerle çalışmak için kullanılmaktadır. Bunun için metinde bulunan semboller ve rakamlar gibi istenmeyen karakterlerin kaldırılması noktasında başvurulmaktadır.

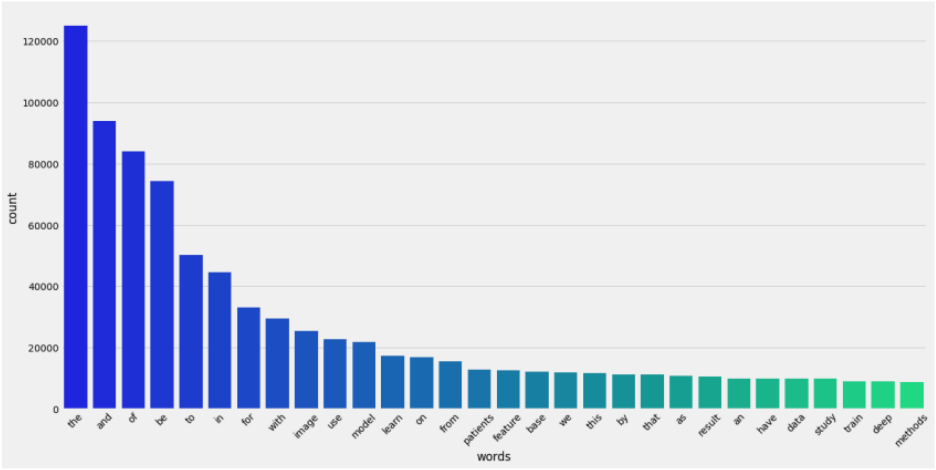
Semboller				;	>	<	£
.	,	?	!	}	[	]	|
#	$	½	§	/	(	)	‘
'	^	+	%	”	\r	\n	“
’	=	\	@	:	{	&	

Şekil 3.3 Veri Önişleme Aşamasında Kaldırılan Semboller

Bu işlemlerden sonra kelimelerin köklerinin bulunması için tokenizasyon ve lemmatizasyon aşamalarına geçilmiştir. Tokenizasyon, metindeki kelimeleri belirlemek için kullanılan bir süreç olurken lemmatizasyon ise aynı kelimeleri kümelemek ve hesaplamak için bir kullanılır.

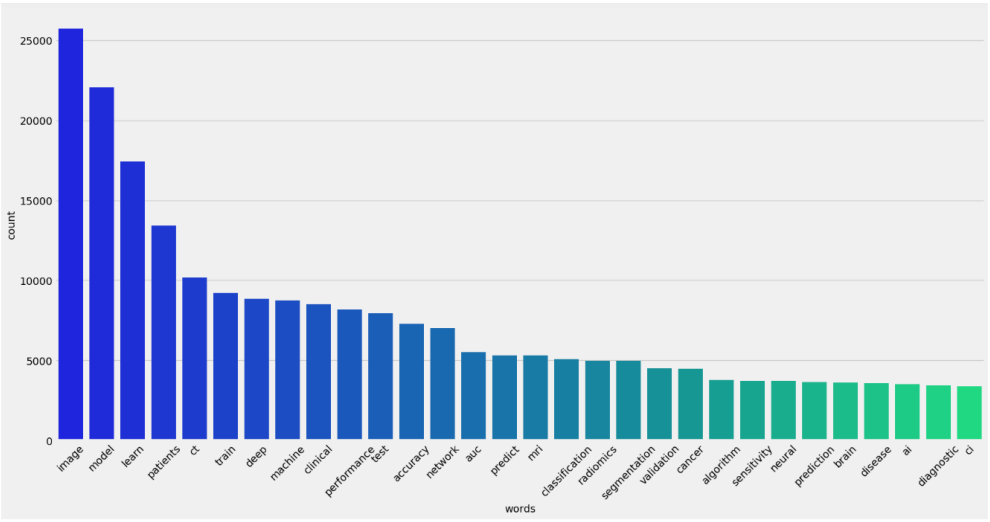
Ayrıca lemmatizasyon, benzer ve gereksiz kelimeleri tespit etmek için de kullanılmaktadır (Foozy ve diğ, 2017). Temizlenen veri "Papers Clean" adında yeni bir dataframe'e aktarılmıştır.

Burada en fazla kullanılan durak kelimeleri (stopwords) görmek için bir histogram grafiği oluşturulmuştur:



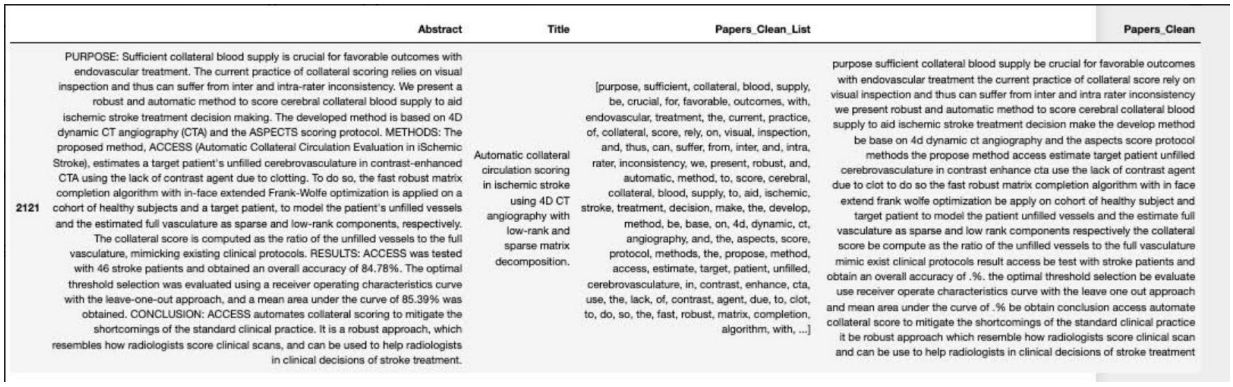
Şekil 3.4 Metinde En Fazla Kullanılan Kelimeler ile Histogram Grafiği (durak kelimeler ile)

Daha sonra durak kelimeler (stopwords) listelenmiştir. Metinde en fazla kullanılan kelimelerden stopwords listesinde olmayanlar bu listeye eklenerek metinden kaldırılmıştır. Daha sonra tekrar metinde en fazla geçen kelimeler ile yeni bir histogram grafiği elde edilmiştir:



Şekil 3.5 Metinde En Fazla Kullanılan Kelimeler ile Histogram Grafiği (durak kelimeler kaldırılarak)

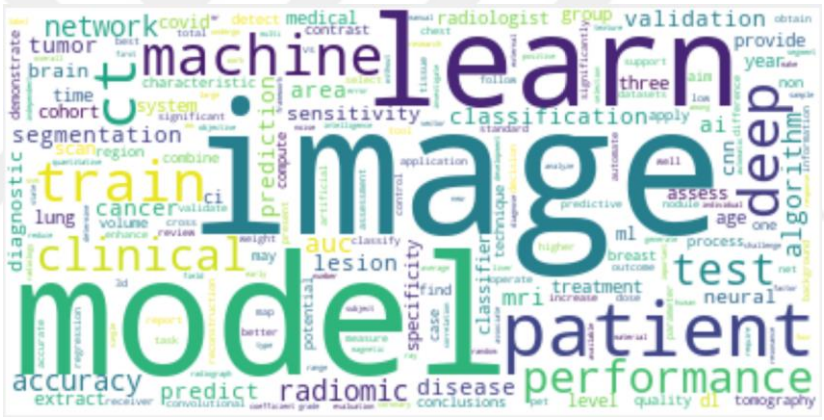
Şekil 3.6'da önişleme sonucu Papers_Clean sütununda temizlenerek analize hazırlanmış özet metinler verilmiştir:



	Abstract	Title	Papers_Clean_List	Papers_Clean
2121	PURPOSE: Sufficient collateral blood supply is crucial for favorable outcomes with endovascular treatment. The current practice of collateral scoring relies on visual inspection and thus can suffer from inter and intra-rater inconsistency. We present a robust and automatic method to score cerebral collateral blood supply to aid ischemic stroke treatment decision making. The developed method is based on 4D dynamic CT angiography (CTA) and the ASPECTS scoring protocol. METHODS: The proposed method, ACCESS (Automatic Collateral Circulation Evaluation in iSchemic Stroke), estimates a target patient's unfilled cerebrovasculature in contrast-enhanced CTA using the lack of contrast agent due to clotting. To do so, the fast robust matrix completion algorithm with in-face extended Frank-Wolfe optimization is applied on a cohort of healthy subjects and a target patient, to model the patient's unfilled vessels and the estimated full vasculature as sparse and low-rank components, respectively. The collateral score is computed as the ratio of the unfilled vessels to the full vasculature, mimicking existing clinical protocols. RESULTS: ACCESS was tested with 46 stroke patients and obtained an overall accuracy of 84.78%. The optimal threshold selection was evaluated using a receiver operating characteristics curve with the leave-one-out approach, and a mean area under the curve of 85.39% was obtained. CONCLUSION: ACCESS automates collateral scoring to mitigate the shortcomings of the standard clinical practice. It is a robust approach, which resembles how radiologists score clinical scans, and can be used to help radiologists in clinical decisions of stroke treatment.	Automatic collateral circulation scoring in ischemic stroke using 4D CT angiography with low-rank and sparse matrix decomposition.	[purpose, sufficient, collateral, blood, supply, be, crucial, for, favorable, outcomes, with, endovascular, treatment, the, current, practice, of, collateral, score, rely, on, visual, inspection, and, thus, can, suffer, from, inter, and, intra, rater, inconsistency, we, present, robust, and, automatic, method, to, score, cerebral, collateral, blood, supply, to, aid, ischemic, stroke, treatment, decision, make, the, develop, method, be, base, on, 4d, dynamic, ct, angiography, and, the, aspects, score, protocol, methods, the, propose, method, access, estimate, target, patient, unfilled, cerebrovasculature, in, contrast, enhance, cta, use, the, lack, of, contrast, agent, due, to, clot, to, do, so, the, fast, robust, matrix, completion, algorithm, with, ...]	purpose sufficient collateral blood supply be crucial for favorable outcomes with endovascular treatment the current practice of collateral score rely on visual inspection and thus can suffer from inter and intra rater inconsistency we present robust and automatic method to score cerebral collateral blood supply to aid ischemic stroke treatment decision make the develop method be base on 4d dynamic ct angiography and the aspects score protocol methods the propose method access estimate target patient unfilled cerebrovasculature in contrast enhance cta use the lack of contrast agent due to clot to do so the fast robust matrix completion algorithm with in face extend frank wolfe optimization be apply on cohort of healthy subject and target patient to model the patient unfilled vessels and the estimate full vasculature as sparse and low rank components respectively the collateral score be compute as the ratio of the unfilled vessels to the full vasculature mimic exist clinical protocols result access be test with stroke patients and obtain an overall accuracy of .%. the optimal threshold selection be evaluate use receiver operate characteristics curve with the leave one out approach and mean area under the curve of .% be obtain conclusion access automate collateral score to mitigate the shortcomings of the standard clinical practice it be robust approach which resemble how radiologists score clinical scan and can be use to help radiologists in clinical decisions of stroke treatment

Şekil 3.6 Önişleme Sonucu Elde Edilen Özet Metinler

Word Cloud kelime methodu doğal dil işlemede konu modelleme çalışmalarında sıklıkla kullanılan görselleştirme yöntemlerinden biridir. Bu bağlamda özetlerde en fazla geçen kelimeler ile bir kelime bulutu (Word Cloud) oluşturulmuştur. Burada da görüleceği gibi "image, model, machine learning, ct, radiomic" kelimeleri sıklıkla kullanılmıştır



Şekil 3.7 En sık kullanılan kelimeler ile kelime bulutu (Word Cloud)

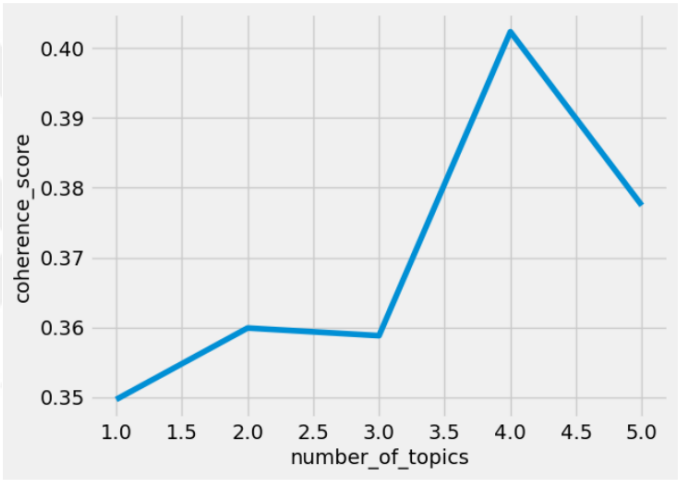
Konu tutarlılığı (Topic Coherence - TC), öğrenilen konuların anlamsal doğasının niteliksel bir değerlendirmesidir ve bir konunun insan yargısına dayalı olarak yorumlanabilirliğini ölçer. Özetle model aracılığıyla üretilen konulardaki kelime topluluğunu gözlemler ve bu gözlemin konunun ne kadar açıklayıcı olduğunu ölçer (Tijare ve Rani, 2020). Konular kullanıcılara tipik olarak en iyi N konu terimleri aracılığıyla sunulduğundan, bir konunun tutarlılığı, en iyi N terimlerinin toplu olarak bir konuyu veya temayı aktarıp aktarmadığına göre değerlendirilir (Chang ve diğ, 2009; Newman ve diğ, 2009; Newman ve diğ, 2010; Lau ve diğ, 2010).

Konu tutarlılığı, model tarafından oluşturulan konulardaki kelimelerin toplanmasını gözlemler ve konunun ne kadar bilgilendirici olduğunu ölçer. Tutarlılık değerini çeşitli şekillerde hesaplayan çok sayıda ölçü vardır. Değerlendirilen konu tutarlılık yöntemleri, konunun en çok

kullanılan kelimelerinin girdisini alır ve kelime çiftleri üzerinden doğrulama ölçütünün toplamını hesaplar.

Tablo 3.1 Konu Tutarlılık Skoru

Konu Sayısı	Coherence Score
1	0.349665
2	0.359906
3	0.358810
4	0.402371
5	0.377477



Şekil 3.8 Coherence Score Grafiği

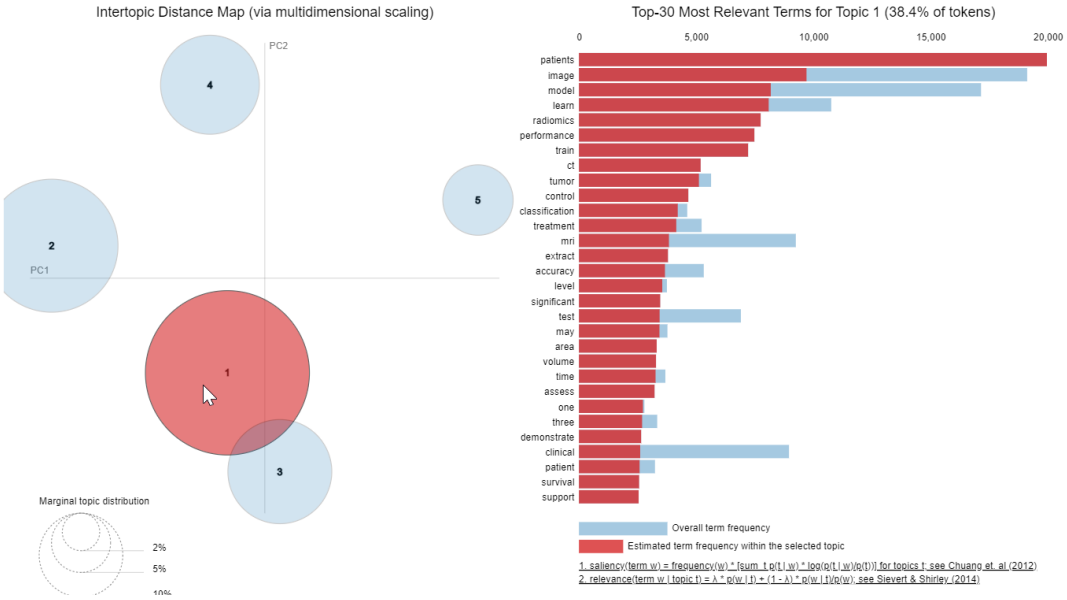
Tablo 3.1 ve Şekil 3.8'den de görülebildiği 2'den fazla konu sayısı ile tutarlılık puanlarının arttığı görülmektedir. Her ne kadar konu sayısı 4 olduğunda en yüksek tutarlılık puanına ulaşılmış olsa da konu sayısı 5 için tutarlılık puanının çok farklı olmaması ve elde edilen sonuçların daha tatmin edici olması sebebiyle konu sayısı 5 olarak seçilmiştir.

Konu Görselleştirme

Konu sayısı seçildikten sonra her konudaki en sık kullanılan sözcükleri, o konunun altında yatan konunun temsili olarak kaydettik. Daha sonra, her konu için merkezi kavramları temsil etmek üzere kelime bulutları dahil edilmiştir. Yazı tipi boyutları, bir kelimenin bir konunun içindeki sıklığıyla orantıdır.

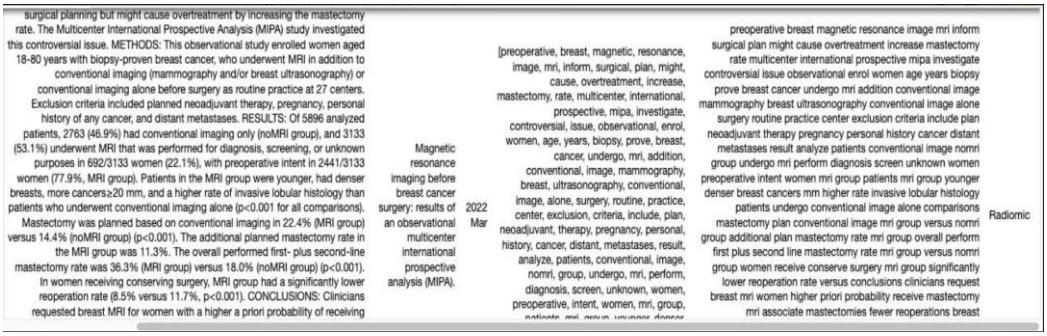
LDA'nın Scikit-learn paketi uygulaması, tutarlılık puanını ölçmek için yöntemler sağlamamaktadır. Uygun konu sayısı dikkate alınarak, LDA modeli Gensim'deki pyLDAvis kullanılarak oluşturulmuş ve görselleştirilmiştir (Tijare ve Rani, 2020).

Gensim paketi ile PyLDAvis modülü, yorumlama kolaylığı için görsel olarak daha etkileşimli bir şekil üretmektedir. Şekil 3.9 - 3.10 - 3.11 - 3.12 – 3.13'te, başlık olasılıklarını temsil eden, konular arası bir mesafe haritası üzerinde çizilen üç konuyu görebiliriz. Sağda veri önişleme ve LDA modellemesinden sonra derlemdeki ilk 30 kelimeyi görüyoruz. Daireler, derlem içindeki belirlenmiş konuları temsil eder ve boyutları, derlem içindeki konu varlığını temsil eder. Daireler arasındaki mesafe, her konu arasındaki benzerliği gösterir (Ki ve diğ, 2021).

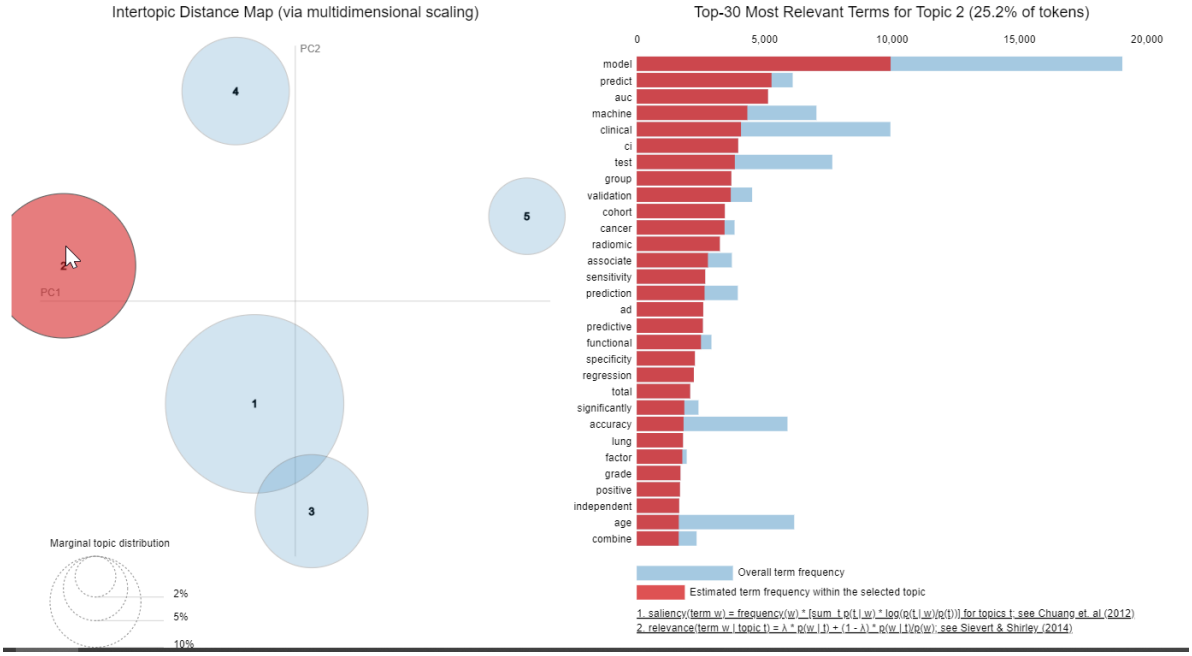


Şekil 3.9 Konu 1 için LDA tarafından tanımlanan konular için pyLDAvis grafiği

Şekil 3.9'da tıbbi görüntülemeden madencilikle elde edilebilen verilerin çıkarılmasını ifade eden radyomiks (radiomic) (Shur ve diğ, 2021) kelimesinin öne çıktığını görüyoruz. Bu yöntem bir görüntü madenciliği aracı olarak görülmekte ve doğal olarak makine öğrenmesi ve derin öğrenme algoritmalarının uygulanmasına katkıda bulunmaktadır (Avanzo ve diğ, 2020). Bu terim makine öğrenmesi ile ilişkili olduğu gibi tıbbın birçok alanında da kullanılmakta olup bunlardan biri onkolojidir ve terimlerin arasında "tumor" kelimesinin de bulunduğu görülmektedir. Görüntüleme ile ilgili bir terim olduğu için "ct" ve "mri" kelimeleri de yine bu kelimelerin arasında yer almaktadır. Ayrıca "train" kelimesinden yola çıkarak uygulamaların denetimli öğrenme ile yapıldığı anlaşılmaktadır.

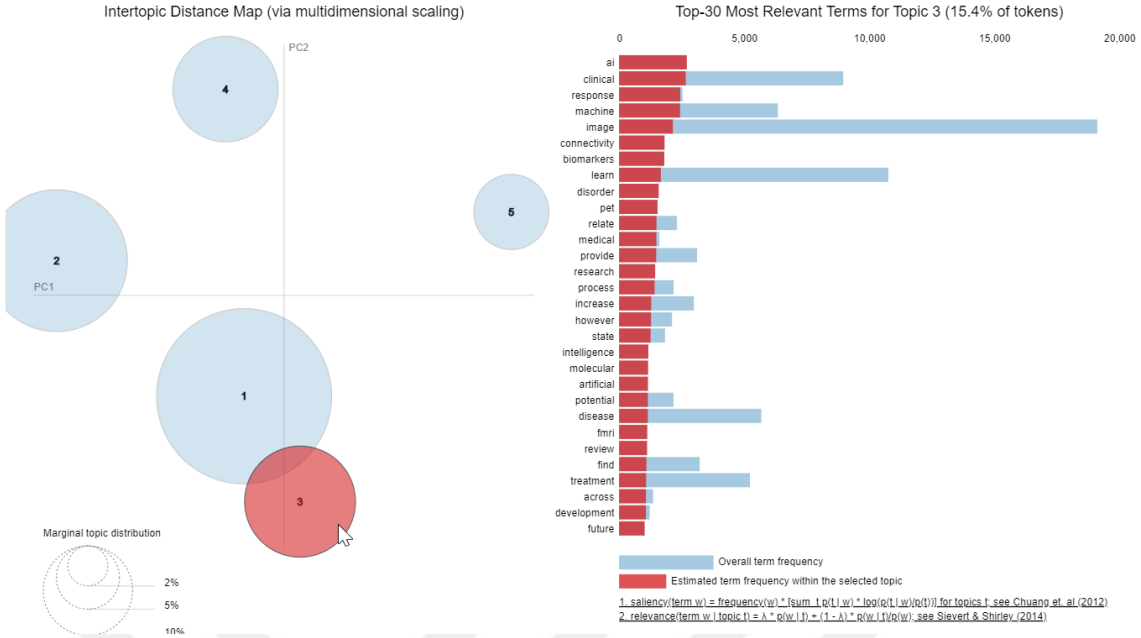


Şekil.3.10 Konu 1'e ait topic örneği



Şekil.3.11 Konu 2 için LDA tarafından tanımlanan konular için pyLDAvis grafiği

Şekil 3.11'de model, tahmin (predict) ve auc gibi makine öğrenmesi yöntemlerine ait kelimelerin öne çıktığı görülmektedir.

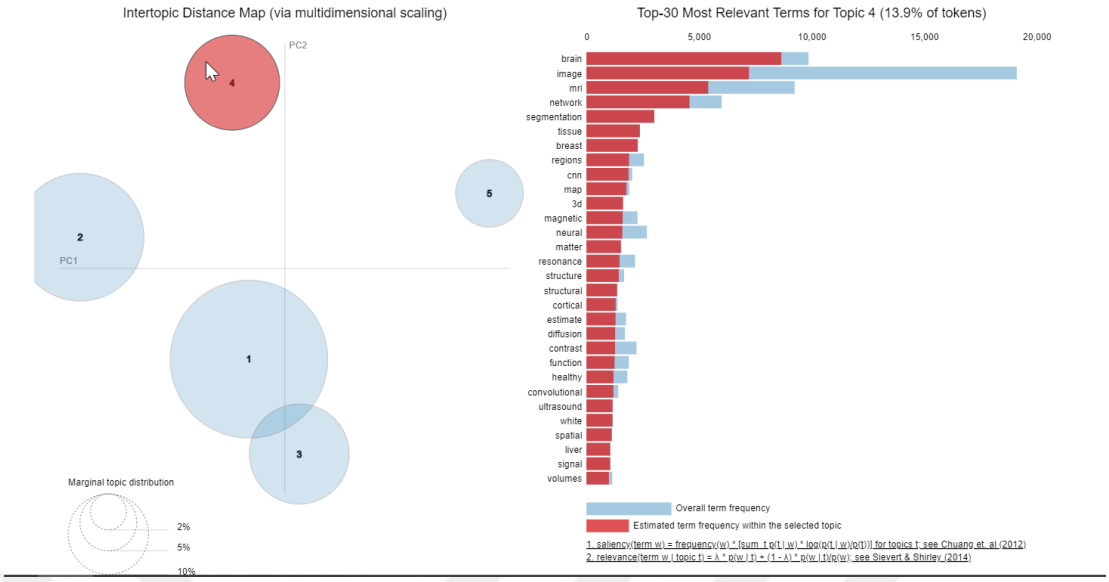


Şekil 3.12 Konu 3 için LDA tarafından tanımlanan konular için pyLDAvis grafiği

Şekil 3.12'de yapay zeka (ai), klinik (clinical) ve makine öğrenmesi (machine learning) kelimeleri öne çıkmış bu da bize bu konunun klinik yapay zeka kapsamında ele alınabildiğini göstermiştir.

BACKGROUND: It is time-consuming for oncologists to delineate volumes for adiotherapy treatment in computer tomography (CT) images. Automatic delineation based on image processing exists, but with varied accuracy and moderate time savings. Using convolutional neural network (CNN), delineations of volumes are faster and more accurate. We have used CTs with the annotated structure sets to train and evaluate a CNN. MATERIAL AND METHODS: The CNN is a standard segmentation network modified to minimize memory usage. We used CTs and structure sets from 75 cervical cancers and 191 anorectal cancers receiving radiation therapy at Skåne University Hospital 2014-2018. Five structures were nvestigated: left/right femoral heads, bladder, bowel bag, and clinical target volume of lymph nodes (CTVNs). Dice score and mean surface distance (MSD) (mm) valuated accuracy, and one oncologist qualitatively evaluated auto-segmentations. RESULTS: Median Dice/MSD scores for anorectal cancer: 0.91-0.92/1.93-1.86 emoral heads, 0.94/2.07 bladder, and 0.83/6.80 bowel bag. Median Dice scores for cervical cancer were 0.93-0.94/1.42-1.49 femoral heads, 0.84/3.51 bladder, 0.88/5.80 bowel bag, and 0.82/3.89 CTVNs. With qualitative evaluation, performance on femoral heads and bladder auto-segmentations was mostly excellent, but CTVN auto-segmentations were not acceptable to a larger extent. ISCUSSION: It is possible to train a CNN with high overlap using structure sets as ground truth. Manually delineated pelvic volumes from structure sets do not always strictly follow volume boundaries and are sometimes inaccurately defined, which leads to similar inaccuracies in the CNN output. More data that is consistently annotated is needed to achieve higher CNN accuracy and to enable future clinical	Auto-segmentations by convolutional neural network in cervical and anorectal cancer with clinical structure sets as the ground truth.	2020 Nov	[background, time, consume, oncologists, delineate, volumes, radiotherapy, treatment, computer, tomography, ct, image, automatic, delineation, image, process, exist, vary, accuracy, moderate, time, save, convolutional, neural, network, cnn, delineations, volumes, faster, accurate, cts, annotate, structure, train, cnn, material, cnn, standard, segmentation, network, modify, minimize, memory, usage, cts, structure, cervical, cancers, anorectal, cancers, receive, radiation, therapy, skåne, university, hospital, five, structure, investigate, leave, right, femoral, head, bladder, bowel, bag, clinical, target, volume, lymph, nod, ctvns, dice, surface, distance, msd, mm, accuracy, one, oncologist, qualitatively, auto, segmentations, median, dice, msd, anorectal, cancer, femoral, head, bladder, bowel, bag, median, dice, cervical, cancer, femoral, head, bladder, ...]	background time consume oncologists delineate volumes radiotherapy treatment computer tomography ct image automatic delineation image process exist vary accuracy moderate time save convolutional neural network cnn delineations volumes faster accurate cts annotate structure train cnn material cnn standard segmentation network modify minimize memory usage cts structure cervical cancers anorectal cancers receive radiation therapy skåne university hospital five structure investigate leave right femoral head bladder bowel bag clinical target volume lymph nod ctvns dice surface distance msd mm accuracy one oncologist qualitatively auto segmentations median dice msd anorectal cancer femoral head bladder bowel bag median dice cervical cancer femoral head bladder bowel bag ctvns qualitative evaluation performance femoral head bladder auto segmentations mostly excellent ctvn auto segmentations acceptable larger extent discussion possible train cnn overlap structure grind truth manually delineate pelvic volumes structure always strictly follow volume boundaries sometimes inaccurately define lead similar inaccuracies cnn output consistently annotate need higher cnn accuracy enable future clinical implementation	Klinik Yapay Zeka

Şekil 3.13 Konu 3'e ait topic örneği

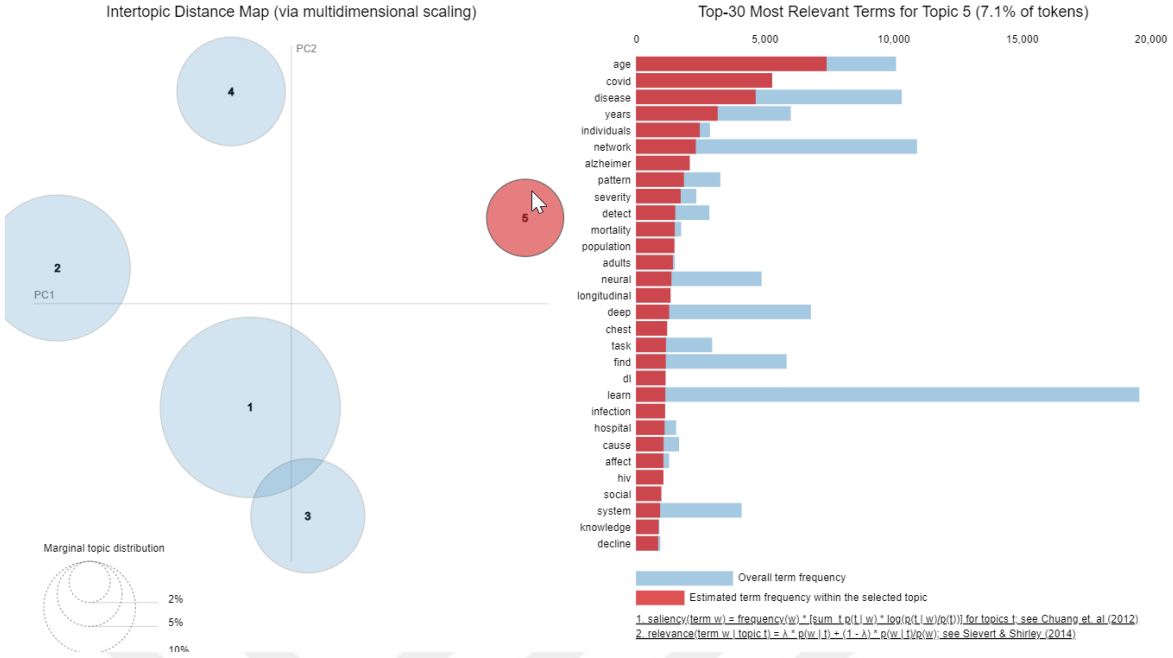


Şekil 3.14 Konu 4 için LDA tarafından tanımlanan konular için pyLDAvis grafiği

Şekil 3.14'te beyin (brain), göğüs (breast), doku (tissue) gibi kanseri konu alan kelimeler ve manyetik rezonans (magnetic resonance), harita (map), 3d ve Evrişimli Sinir Ağları (Convolutional Neural Network (CNN)) gibi görüntüleme konularını ele alan kelimeler karşımıza çıkmaktadır. Bu yöntem radyoloji dahil olmak üzere bilgisayarlı görü üzerine çalışan farklı alanlarda kullanılmaktadır (Yamashita ve diğ, 2018).

	Abstract	Title	Year	Papers_Clean_List	Papers_Clean	Topic
	Machine learning (ML) is a burgeoning field of medicine with huge resources being applied to fuse computer science and statistics to medical problems. Proponents of ML extol its ability to deal with large, complex and disparate data, often found within medicine and feel that ML is the future for biomedical research, personalized medicine, computer-aided diagnosis to significantly advance global health care. However, the concepts of ML are unfamiliar to many medical professionals and there is untapped potential in the use of ML as a research tool. In this article, we provide an overview of the theory behind ML, explore the common ML algorithms used in medicine including their pitfalls and discuss the potential future of ML in medicine.	eDoctor: machine learning and the future of medicine.	2018 Dec	[machine, learn, ml, burgeon, field, medicine, huge, resources, apply, fuse, computer, science, statistics, medical, problems, proponents, ml, extol, ability, deal, large, complex, disparate, often, find, within, medicine, feel, ml, future, biomedical, research, personalize, medicine, computer, aid, significantly, advance, global, health, care, however, concepts, ml, unfamiliar, many, medical, professionals, untapped, potential, ml, research, tool, article, provide, overview, theory, behind, ml, explore, common, ml, algorithms, medicine, pitfalls, discuss, potential, future, ml, medicine]	machine learn ml burgeon field medicine huge resources apply fuse computer science statistics medical problems proponents ml extol ability deal large complex disparate often find within medicine feel ml future biomedical research personalize medicine computer aid significantly advance global health care however concepts ml unfamiliar many medical professionals untapped potential ml research tool article provide overview theory behind ml explore common ml algorithms medicine pitfalls discuss potential future ml medicine	Kanser ve Görüntüleme

Şekil 3.15 Konu 4'e ait topic örneği



Şekil 3.16 Konu 5 için LDA tarafından tanımlanan konular için pyLDAvis grafiği

Şekil 3.16'da hastalık (disease), koronavirüs (Covid-19), Alzheimer, göğüs (chest) gibi hastalıklarla ilgili kelimelerin yanı sıra yaş (age) ve yıl (years) bu konunun tıbbi hastalıklar kapsamına girdiği anlaşılmaktadır.

Infarct volume (FIV) on follow-up diffusion-weighted imaging (FU-DWI) is only moderately associated with functional outcome in acute ischemic stroke patients. However, FU-DWI may contain other imaging biomarkers that could aid in improving outcome prediction models for acute ischemic stroke. We included FU-DWI data from the HERMES, ISLES, and MR CLEAN-NO IV databases. Lesions were segmented using a deep learning model trained on the HERMES and ISLES datasets. We assessed the performance of three classifiers in predicting functional independence for the MR CLEAN-NO IV trial cohort based on: (1) FIV alone, (2) the most important features obtained from a trained convolutional autoencoder (CAE), and (3) radiomics. Furthermore, we investigated feature importance in the radiomic-feature-based model. For outcome prediction, we included 206 patients: 144 scans were included in the training set, 21 in the validation set, and 41 in the test set. The classifiers that included the CAE and the radiomic features showed AUC values of 0.88 and 0.81, respectively, while the model based on FIV had an AUC of 0.79. This difference was not found to be statistically significant. Feature importance results showed that lesion intensity heterogeneity received more weight than lesion volume in outcome prediction. This study suggests that predictions of functional outcome should not be based on FIV alone and that FU-DWI images capture additional prognostic information.	Outcome Prediction Based on Automatically Extracted Infarct Core Image Features in Patients with Acute Ischemic Stroke.	2022 Jul 23	[infarct, volume, fiv, follow, diffusion, weight, image, fu, dwi, moderately, associate, functional, outcome, acute, ischemic, stroke, patients, however, fu, dwi, may, contain, image, biomarkers, could, aid, outcome, prediction, model, acute, ischemic, stroke, fu, dwi, hermes, isles, mr, clean, iv, databases, lesions, segment, deep, learn, model, train, hermes, isles, datasets, assess, performance, three, classifiers, predict, functional, independence, mr, clean, iv, trial, cohort, fiv, alone, important, obtain, train, convolutional, autoencoder, cae, radiomics, furthermore, investigate, importance, radiomic, model, outcome, prediction, patients, scan, train, validation, test, classifiers, cae, radiomic, auc, model, fiv, auc, difference, find, statistically, significant, importance, lesion, intensity, heterogeneity, receive, weight, lesion, ...]	infarct volume fiv follow diffusion weight image fu dwi moderately associate functional outcome acute ischemic stroke patients however fu dwi may contain image biomarkers could aid outcome prediction model acute ischemic stroke fu dwi hermes isles mr clean iv databases lesions segment deep learn model train hermes isles datasets assess performance three classifiers predict functional independence mr clean iv trial cohort fiv alone important obtain train convolutional autoencoder cae radiomics furthermore investigate importance radiomic model outcome prediction patients scan train validation test classifiers cae radiomic auc model fiv auc difference find statistically significant importance lesion intensity heterogeneity receive weight lesion volume outcome prediction suggest predictions functional outcome fiv alone fu dwi image capture additional prognostic information	Tıbbi Hastalıklar

Şekil 3.17 Konu 5'e ait topic örneği

Her 5 konuda makine öğrenmesi, derin öğrenme ve görüntüleme ile ilgili terimlere ait kelimeler yer almaktadır. Buradan yola çıkarak makine öğrenmesi yöntemlerinin görüntüleme alanında kullanıldığı ifade edilebilmektedir.

Birinci, ikinci, üçüncü, dördüncü ve beşinci konular, en yüksek frekanslara sahip ilk 10 terim şu şekilde gösterilebilmektedir:

```
Topic: 0 Word: 0.041*"age" + 0.029*"covid" + 0.026*"disease" + 0.018*"y
ears" + 0.014*"individuals" + 0.013*"network" + 0.012*"alzheimer" + 0.0
10*"pattern" + 0.010*"severity" + 0.008*"detect"
```

```
Topic: 1 Word: 0.037*"patients" + 0.018*"image" + 0.015*"model" + 0.015
*"learn" + 0.014*"radiomics" + 0.014*"performance" + 0.013*"train" + 0.
010*"ct" + 0.009*"tumor" + 0.009*"control"


Topic: 2 Word: 0.044*"brain" + 0.037*"image" + 0.028*"mri" + 0.023*"net
work" + 0.015*"segmentation" + 0.012*"tissue" + 0.012*"breast" + 0.010*
"regions" + 0.010*"cnn" + 0.009*"map"


Topic: 3 Word: 0.012*"ai" + 0.012*"clinical" + 0.011*"response" + 0.011
*"machine" + 0.010*"image" + 0.008*"connectivity" + 0.008*"biomarkers"
+ 0.008*"learn" + 0.007*"disorder" + 0.007*"pet"


Topic: 4 Word: 0.025*"model" + 0.013*"predict" + 0.013*"auc" + 0.011*"m
achine" + 0.010*"clinical" + 0.010*"ci" + 0.010*"test" + 0.009*"group"
+ 0.009*"validation" + 0.009*"cohort"
```

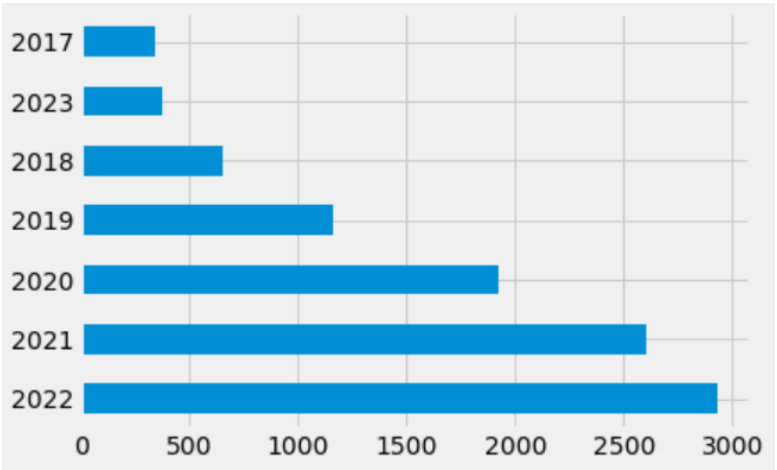
Burada birinci konu (Topic 0) tıbbi hastalıklar konusunu, ikinci konu (Topic 1) radyomiks (radiomics) konusunu, üçüncü konu (Topic 2) kanser ve görüntüleme konusunu, dördüncü konu (Topic 3) klinik yapay zeka ve beşinci konu (Topic 4) ise makine öğrenmesi (sınıflandırma) konusunu temsil eden terimleri kapsamaktadır.

4. BULGULAR

Tablo 4.1'de 1995-2023 yılları arasında yapılan makine öğrenmesi ve radyoloji ile ilgili çalışmaların sayısı yer almaktadır. Tablo 4.1'den yola çıkılarak elde edilen Şekil 4.1'den de anlaşılabileceği üzere makine öğrenmesi ve radyoloji ile ilgili yapılan çalışmaların üstel olarak arttığı görülmektedir. 2022 yılında biraz daha düşmüş olsa da özellikle son 4 yılda çalışmalar katlanarak artmıştır.

Tablo 4.1 PubMed'de Yıllara Göre Yayımlanan Çalışma Sayıları

Yıl	Çalışma Sayısı	Yıl	Çalışma Sayısı	Yıl	Çalışma Sayısı
1995	1	2007	12	2018	842
1996	1	2008	17	2019	1458
1997	2	2009	15	2020	2313
1998	2	2010	21	2021	3045
1999	1	2011	36	2022	2895
2000	2	2012	76	2023	379
2001	6	2013	105		
2002	2	2014	148		
2004	3	2015	239		
2005	9	2016	297		
2006	6	2017	453		



Şekil 4.1 PubMed'de Yıllara Göre Yayımlanan Çalışma Sayıları (Veri Önişleme Yapıldıktan Sonra Elde Edilen Özetlerin Dağılımı)

Tez çalışmasında radyolojide makine öğrenmesi yöntemlerinden yararlanılan makale çalışmalarının özetleri üzerinden doğal dil işleme yöntemleri kullanılarak bir bibliyografik çalışma gerçekleştirilmiştir. Çalışmanın sonucunda LDA yöntemi ile 9997 makale özeti

incelenmiştir. Elde edilen bulgulara göre makale özetlerinin tıbbi hastalıklar, radiomics, kanser ve görüntüleme, klinik yapay zeka ve makine öğrenmesi (sınıflandırma) olmak üzere 5 ana konu başlığı altında toplandığı görülmektedir. Covid ve Alzheimer gibi son yılların önemli hastalıkları üzerine makine öğrenmesi çalışmalarının arttığı ve aynı konu altında toplandığı görülmüştür. Bir diğer konu olan görüntüden veri çıkarmaya karşılık gelen raidomicsin de çalışmaların en büyük bölümünü oluşturduğu görülmüştür. Makine öğrenmesinde sınıflandırma konusunun da en popüler artış trendine sahip çalışma alanının olduğu görülmüştür. Klinik çalışmalar alanında da yapay zeka çalışmalarının bir diğer trend konusu olduğu söylenebilir. Son olarak da kanser ve görüntüleme konusu önemli bir makine öğrenmesi-radyoloji konu başlığı olarak karşımıza çıkmıştır.

5. TARTIŞMA VE SONUÇ

Metin verileri genel olarak yapılandırılmamış veri kaynaklarını oluşturmaktadır. Bu verilerin işlenerek bir öngörü elde edilmesi için makine diline çevrilerek işlenmesi gerekmektedir. Bu aşamada devreye Doğal Dil İşleme yöntemleri girmektedir. Doğal Dil İşleme, insanların konuştuğu dillerin makine diline aktarılmasını sağlayan bir yapay zeka alt alanıdır.

Sağlık alanında her geçen gün artan metin verileri (hasta taburcu raporları, görüntü raporları vb.) yapılandırılmamış veri kaynaklarını oluşturmaktadır. Bu yapılandırılmamış veri kaynakları birtakım istatistiksel ve yapay öğrenme yöntemlerinin yanı sıra kural tabanlı ve algoritmik yaklaşımlarla makine diline çevrilmektedir ve bu sayede bu veri kaynaklarından anlamlı sonuçlar elde edilmektedir.

Makine öğrenmesi algoritmalarının temel amacı veriden öngörü elde etmektir. Günümüzde ileri makine öğrenmesi algoritmaları ile yapay zeka alanında gündelik işlerin otomizasyonunun maksimize edilerek hataların minimize edilmesi amaçlanır. Doğal dil işlemede de insanın uzun sürede yapabileceği metin çözümlemelerinin otomatize edilerek daha kısa sürelerde daha fazla metinden öngörü elde etmesi hedeflenir. Buna paralel olarak görüntü işleme gibi farklı makine öğrenmesi alanında da görüntü hammaddesinden otomatize edilecek öngörüler elde etmek istenir. Bu bağlamda radyolojinin son yıllarda görüntü işleme başta olmak üzere makine öğrenmesi yöntemlerine sıklıkla başvurduğunu görmekteyiz. Bu sebeple çalışmada konunun popülerliğinden dolayı radyoloji ve makine öğrenmesi olarak seçilmiş ve trendin ne yönde gelişeceği incelenmiştir.

Bibliyografik çalışmalarda R programlama dili başta olmak üzere çok fazla bibliyografi kütüphanesi bulunmaktadır. Bunlar ile yazar bilgileri, kurum bilgileri, çalışma başlığı, anahtar kelimeler gibi bilgiler indirgenip otomatik olarak tanımlayıcı istatistikler elde edilebilmektedir. Bu tez çalışmasında bu klasik bibliyografik özet çalışmalarından farklı olarak çalışma özetlerinden yola çıkarak anlamsal yönelimin keşfedilmesi amaç edinilmiştir. Bundan dolayı da bilgi çıkarım yöntemlerinden konu modelleme ile bibliyografik bir çalışma gerçekleştirilmiştir.

Bulgularda da detaylı bir şekilde aktarıldığı gibi makale özetleri tıbbi hastalıklar, radiomics, kanser ve görüntüleme, klinik yapay zeka ve makine öğrenmesi (sınıflandırma) olmak üzere 5 ana konu başlığı altında toplanmıştır.

Literatürde incelenen ilaç sektöründen eğitime acil tıptan pazarlamaya kendine çalışma alanı bulan doğal dil çalışmalarından da anlaşılabildiği gibi makale özetlerinden trend analizi çıkaran çalışmalarda konu modelleme ve LDA kullanılmıştır. Veri ön işlemede de köklendirme ve lematizasyon önemli bir yer tutmuştur.

LDA yöntemi konu modelleme için sıklıkla kullanılan bir yöntem olmakla birlikte özellikle kelimeler arası bağlılıkları ortaya çıkarmada bazı durumlarda yetersiz kalabilmektedir. Bundan dolayı n-gram gibi kelimeler arası ilişki kuran ileri doğal dil işleme yöntemlerinin kullanılması gerekebilir. Bu tez çalışmasında ana odak konusu literatürün yöneldiği konuların hangi alanlarda toplandığını gösterebilecek bir trend analizi çıkarımı yapmak olup bu doğrultuda da LDA ile sonuca ulaşılmıştır.

Çalışmanın sonucuna dayanarak, Konu Modelleme yöntemlerinin radyolojide makine öğrenmesi konusunda uygulanma potansiyeli olduğu ortaya konmuştur.

KAYNAKLAR

Abdelaziz, A., Santos, V., Dias, M. S., 2021, Machine Learning Techniques in the Energy Consumption of Buildings: A Systematic Literature Review Using Text Mining and Bibliometric Analysis, *Energies*, 14, 7810.

Allam, A. M. N., Haggag, M. H., 2012, The Question Answering Systems: A Survey, *International Journal of Research and Reviews in Information Sciences (IJRRIS), 2*(3).

Amado, A., Cortez, P., Rita, P., Moro, S., 2017, Research Trends on Big Data in Marketing: A Text Mining and Topic Modeling Based Literature Analysis, *European Research on Management and Business Economics*, 24(2018), 1-7.

Audrin, C., Audrin, B., 2021, Key Factors in Digital Literacy in Learning and Education: A Systematic Literature Review Using Text Mining, *Education and Information Technologies*, 27(2022), 7395–7419.

Avanzo M., Wei, L., Stancanello, J., Vallières, M., Rao, A., Morin, O., Mattonen, S.A., El Naqa, I., 2020, Machine and Deep Learning Methods for Radiomics. *Medical Physics*.

Beam, A. L., Kohane, I. S., 2018, Big Data and Machine Learning in Health Care, *Jama*, 319(13), 1317-1318.

Bird, S., Klein, E., Loper, E., 2009, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, Inc.

Blei, D. M., 2011, Introduction to Probabilistic Topic Models. *Communications of the ACM*.

Brill, E., Mooney, R. J., 1997, An Overview of Empirical Natural Language Processing. *AI magazine*, *18*(4), 13-13.

Brynjolfsson, E., Mcafee, A., 2017, Artificial Intelligence for Real. *Harvard Business Review*, 1, 1-31.

Cai, T., Giannopoulos, A. A., Yu, S, Kelil, T., Ripley, B., Kumamaru, K. K., Rybicki, F. J., Mitsouras, D., 2016, Natural Language Processing Technologies in Radiology Research and Clinical Applications, *Radiographics*, 36(1), 176-91.

Cai, T., Giannopoulos, A. A., Yu, S., Kelil, T., Ripley, B., Kumamaru, K. K., Rybicki, F. J., Mitsouras, D., 2016, Natural Language Processing Technologies in Radiology Research and Clinical Application, *RadioGraphics,* 36(1), 177-191.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D., 2009, Reading Tea Leaves: How Humans Interpret Topic Models. *In Proceedings of Twenty-Fourth Annual Conference on Neural Information Processing Systems* (NIPS 2009). 288–296.

Chen, X., Zou, D., Xie, H., Su, F., 2021, Twenty-five Years of Computer-Assisted Language Learning: A Topic Modeling Analysis, *Language Learning & Technology*, 25(3), 151-185.

Choo, J., Lee, C., Reddy, C. K., Park, H., 2013, Utopian: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization, *IEEE transactions on visualization and computer graphics*, *19*(12), 1992-2001.

Ki, C.N., Hosseinian-Far, A., Daneshkhah, A., Salari, N., 2021, Topic Modellig in Precisin Medicine with Its Applications in Personalized Diabets Management, *Expert Systems*.

Chopra, A., Prashar, A., Sain, C., 2013, Natural Language Processing, *International Journal of Technology Enhancements and Emerging Engineering Research*, 1(4), 131-134

Dangeti, P., 2017, *Statistics for Machine Learning.* Packt Publishing Ltd.

David M. B., Andrew Y. N., Michael I. J., 2003, Latent Dirichlet Allocation *Journal of Machine Learning Research,* 3, 993–1022.

Dharwadkar, R., Deshpande, N. A., 2018, A Medical Chatbot. *International Journal of Computer Trends and Technology (IJCTT)*, *60*(1), 41-45.

Doğuç, O, 2022, Twitter Verisi İle Doğal Afet Müdahale Süreci İçin Karar Destek Uygulaması. Afet ve Risk Dergisi, 5(2), 408-419.

Dumais, S. T., 2004, Latent Semantic Analysis, *Annu. Rev. Inf. Sci. Technol.*, *38*(1), 188-230.

Foozy, C. F. M., Ahmad, R., Abdollah, M. F., Wen, C. C., 2017, A Comparative Study with RapidMiner and WEKA Tools over some Classification Techniques for SMS Spam. *In IOP Conference Series: Materials Science and Engineering*, IOP Publishing, 226(1).

Goel, A. K., DiLella, D., Dotsikas, G., Hilts, M., Kwan, D., Paxton, L., 2019, Unlocking Radiology Reporting Data: an Implementation of Synoptic Radiology Reporting in Low-Dose CT Cancer Screening. *J Digit Imaging*, 32(6), 1044–51.

Griffiths T., Steyvers, M., 2004, Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, 1, 5228–5335.

Guan, R., Wen, X., Liang, Y., Xu, D., He, B., Feng, X., 2019, Trends in Alzheimer's Disease Research Based upon Machine Learning Analysis of PubMed Abstracts, *International Journal of Biological Sciences,* 15(10), 2065-2074.

Hirschman, L., Gaizauskas, R., 2001, Natural Language Question Answering: the View from Here, *Natural Language Engineering*, 7(4), 275-300.

Hutchins, W.J., 1986, *Machine Translation: Past, Present, Furute*, Chichester: Ellis Horwood Limited.

Jay, K. S., Krishnaraj. A., 2017, Strategies for Improving the Value of the Radiology Report: A Retrospective Analysis of Errors in Formally Over-read studies. *J Am Coll Radiol*, 14(4), 459–66.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019, Latent Dirichlet Allocation (LDA) and Topic Modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169-15211.

Jiang, H., Qiang, M., Lin, P., 2015, A Topic Modeling Based Bibliometric Exploration of Hydropower Research, *Renewable and Sustainable Energy Reviews*, 57(2016), 226–237.

Jiang, K., Lu, X., 2020, Natural Language Processing and Its Applications in Machine Translation: A Diachronic Review, *2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI)*, IEEE, 210-214.

Jones, C. L., Bridges, R. A., Huffer, K. M., Goodall, J. R., 2015, Towards a Relation Extraction Framework for Cyber-Security Concepts, *Proceedings of the 10th Annual Cyber and Information Security Research Conference*, 1-4,

Kane, D. A., Rogé, P., Snapp, S. S., 2016, A Systematic Review of Perennial Staple Crops Literature Using Topic Modeling and Bibliometric Analysis, *PLoS ONE*, 11(5).

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., Gurusamy, V., 2014, Preprocessing techniques for text mining. International Journal of Computer Science & Communication Networks, 5(1), 7-16.

Karami, A., Lundy, M., Webb, F., Dwivedi, Y. K., 2020, Twitter and Research: A Systematic Literature Review Through Text Mining, *IEEE*, 8, 67698-67717.

Khurana, D., Koli, A., Khatter, K., Singh, S., 2017, Natural Language Processing: State of the Art, Current Trends and Challenges, *arXiv preprint arXiv:1708.05148*.

Kılıçarslan, Z., Altuğ, E., 2016, Bibliyografya, *Türk Kütüphaneciliği,* 30(2), 338-339.

Koltcov, S., Koltsova, O., Nikolenko, S., 2014, Latent Dirichlet Allocation: Stability and Applications to Studies of User-Generated Content, *Proceedings of the 2014 ACM conference on Web science*, 161-165.

Kulkarni, S. S., Apte, U. M., Evangelopoulos, N. E., 2014, The Use of Latent Semantic Analysis in Operations Management Research, *Decision Sciences*, *45*(5), 971-994.

Kumar, E., 2013, *Natural Language Processing*. IK International Pvt. Ltd.

Lau, J., Newman, D., Karimi, S., Baldwin, T., 2010, *Best Topic Word Selection for Topic Labelling. In Coling 2010: Posters.* Beijing, China, 605–613.

Li, W., McCallum, A., 2006, Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations, *International Conference on Machine Learning (ICML).*

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., Gómez-Berbís, J. M., 2013, Named Entity Recognition: fallacies, challenges and opportunities, *Computer Standards & Interfaces*, 35(5), 482-489.

McCharty, J., 1989, Artificial Intelligence, Logic and Formalizing Common Sense. In Philosophical Logic and Artificial Intelligence, *Springer,* Dordrecht.

Mendonça, E. A., Haas, J., Shagina, L., Larson, E., Friedman, C., 2005, Extracting Information on Pneumonia in Infants Using Natural Language Processing of Radiology Reports, *Journal of Biomedical Informatics,* 38(2005), 314–321.

Munot, N., Govilkar, S. S., 2014, Comparative Study of Text Summarization Methods. *International Journal of Computer Applications*, *102*(12).

Newman, D., Karimi, S., Cavedon, L., 2009, External Evaluation of Topic Models, *In Proceedings of the Fourteenth Australasian Document Computing Symposium (ADCS 2009).* Sydney, Australia, 11–18.

Newman, D., Lau, J. H., Grieser, K., Baldwin, T., 2010, Automatic Evaluation of Topic Coherence. *In Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010).* Los Angeles, USA, 100– 108

Olthof, A.W., van Ooijen, P. M. A., Cornelissen, L. J., 2021, Deep Learning-Based Natural Language Processing in Radiology: The Impact of Report Complexity, Disease Prevalence, Dataset Size, and Algorithm Type on Model Performance. *J Med Syst,* 45(91).

Pons, E., Foks, K. A., Dippel, D. W. J., Hunink, M. G. M., 2019, Impact of Guidelines for the Management of Minor Head Injury on The Utilization and Diagnostic Yield of Ct Over Two Decades, Using Natural Language Processing İn A Large Dataset. *Eur Radio*l, 29(5), 2632–40.

Porter, M.F., 1980, An Algorithm for Suffix Stripping,Program, 14(3), 130-137.

Porturas, T., Taylor, R. A., 2020, Forty Years of Emergency Medicine Research: Uncovering Research Themes and Trends Through Topic Modeling, *American Journal of Emergency Medicine*, 1-8.

Russell, S. J., Norvig, P., 2021, *Artificial Intelligence: A Modern Approach.* Pearson.

Sahni, V. A., Khorasani, R., 2016, The Actionable Imaging Report. *Abdom Radiol.* 41(3), 429–43.

Shur, J. D., Doran, S. J., Kumar, S., Ap Dafydd, D., Downey, K., O'Connor, J. P., Orton, M. R., 2021, Radiomics in Oncology: a practical guide. *Radiographics,* 41(6), 1717-1732.

Singh, S., 2018, Natural Language Processing for Information Extraction, *arXiv:1807.02383*

Spasic I, Nenadic G, Goran Nenadic. Clinical Text Data in Machine Learning: Systematic Review. JMIR Med Informatics. 2020;8(3).

Sperandeo, R., Messina, G., Iennaco, D., Sessa, F., Russo, V., Polito, R., Monda, V., Monda, M., Messina, A., Mosca, L. L., Mosca, L., Dell'Orco, S., Moretto, E., Gigante, E., Chiacchio, A., Scognamiglio, C., Carotenuto, M., and Maldonato, N. M., 2020, What Does Personality Mean in the Context of Mental Health? A Topic Modeling Approach Based on Abstracts Published in Pubmed Over the Last 5 Years, *Frontiers in Psychiatry,* 10(938), 1-12.

Tijare, P., Rani P. J., 2020, Exploring Popular Topic Models, *Journal of Physics: Conference Series,* 1706 (2020).

Tong, Z., Zhang, H., 2016, A Text Mining Research Based on LDA Topic Modelling, *International conference on computer science, engineering and information technology*, 201-210.

URL 1, https://www.chatbots.org/, [Ziyaret Tarihi: 26 Haziran 2022].

URL 2, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3372692/, [Ziyaret Tarihi: 20 Haziran 2022].

Vangara, R., Skau, E., Chennupati, G., Djidjev, H., Tierney, T., Smith, J. P., Alexandrov, B. S., 2020, Semantic Nonnegative Matrix Factorization with Automatic Model Determination for Topic Modeling. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 328-335.

Velupillai, S., Mowery, D., South, B. R., Kvist, M., Dalianis, H., 2015, Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. *Yearbook of medical informatics*, *24*(01), 183-193.

Wang, S., Summers, R. M., 2012, Machine Learning and Radiology, *Med Image Anal*, 16(5), 933–951.

Xiang, W., Wang, B., 2019, A Survey of Event Extraction from Text. *IEEE Access*, *7*, 173111-173137.

Yamashita, R., Nishio, M., Do, R. K. G., Togashi, K., 2018, Convolutional Neural Networks: an overview and application in radiology. *Insights Into Imaging*, 9, 611-629.

Yim, W. W., Yetişgen, M., Harris, W. P., Kwan, G. B., 2016, Onkolojide doğal dil işleme: bir inceleme. *JAMA Oncol*. 2 (6), 797-804.

Zhang, D., Lee, W., 2003, A Web-based Question Answering System, *Massachusetts Institute of Technology*.

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Selin Şahin
Doğum Yeri	
Doğum Tarihi	Tarih girmek için tıklayın veya dokunun.
Uyruğu	☑ T.C. ❒ Diğer:
E-Posta Adresi	
Web Adresi	

Eğitim Bilgileri	
Lisans	
Üniversite	İstanbul Üniversitesi
Fakülte	İktisat Fakültesi
Bölümü	Ekonometri
Mezuniyet Yılı	2018

Yüksek Lisans	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Enformatik Anabilim Dalı
Programı	Enformatik Programı

Doktora	
Üniversite	
Enstitü Adı	
Anabilim Dalı	Anabilim Dalı Adı
Programı	Program Adı

Makale ve Bildiriler