Avrupa Bilim ve Teknoloji Dergisi
Sayı 21, S. 518-526, Ocak 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 21, pp. 518-526, January 2021

Research Article

Metin Madenciliği ile Tıbbi Tedavi Alanlarının Yakınlıklarının Ölçülmesi

Hasan Kurban*

Siirt Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Siirt, Türkiye, (ORCID: 0000-0003-3142-2866), hakurban@gmail.com

(İlk Geliş Tarihi 29 Kasım 2020 ve Kabul Tarihi 30 Ocak 2021)

(**DOI:** 10.31590/ejosat.833199)

ATIF/REFERENCE: Kurban, H. (2021). Metin Madenciliği ile Tıbbi Tedavi Alanlarının Yakınlıklarının Ölçülmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (21), 518-526.

Öz

Bazı hastalık belirtilerinin birçok tıbbi tedavi alanıyla ilgili olması, hastaların tedavi için randevu alırken zorlanmalarına sebep olmaktadır. Örneğin; karın ağrısı rahatsızlığı bulunan bir hastanın rahatsızlığı dahiliye, hariciye ya da intaniye bölümlerinden herhangi birisiyle ilgisi bulunabilmektedir. Bu çalışmada T.C. Sağlık Bakanlığına bağlı birçok kamu hastanesinin resmî internet sitesinde bulunan ve hastaların belirtilerine göre doğru tıbbi tedavi branşını seçmelerine yardımcı olmak amacıyla kullanılan 13 tıbbi alan ve 204 belirti, metin madenciliği ve veri bilimi teknikleriyle kapsamlı olarak incelenmiştir. Kamu hastanelerinin resmî internet sitelerde kullanılan metnin içeriği baz alınarak tıbbi tedavi alanları arasındaki, yakınlık/uzaklık hesaplanıp, kelime bazlı hastaları randevu alanını belirlerken en çok zorlayan kelimeler ve belirtiler tespit edilmiştir. Kullanılan kelimeler analiz edilirken edat ve bağlaç gibi anlamsız sözcükler göz ardı edilip, hastalık belirtileri üzerinde kelime bulutu (word cloud) oluşturulmuştur. Tıbbi alanların yakınlığını hesaplamak için öncelikle metin içeriği kullanılarak 13 alan için her bir belirtinin var olup olmadığını gösteren 13x186 boyutlu ikili veri (binary data, document matrix) oluşturulmuştur. Daha sonra, bu veri seti üzerinde tıbbi tedavi alanları belirtilere göre aglomeratif hiyerarşik kümeleme algoritmaları (single, complete, average, ward, mcquitty) kullanılarak kümelendirilip metin bazlı alanların birbiri ile yakınlığı tespit edilmiştir. Bu makalenin sonuçlar kısmında hastaları en çok zorlayan kelimeler ve tıbbi alanların metin bazlı yakınlıkları paylaşılmıştır. Elde edilen sonuçlar çerçevesinde kullanılan metnin sağlık uzmanları tarafından tekrar düzenlenmesinin, yanlış tıbbi branşlardan alınan randevu sayısının azaltılmasına katkısı olacaktır.

Anahtar Kelimeler: Hastane randevusu, Metin madenciliği, Veri bilimi, Hastalık belirtileri, Sağlık.

Measuring the Proximity of Medical Treatment Areas with Text Mining

Abstract

The fact that some of the symptoms are related to many medical treatment areas causes patients to have difficulty in making an appointment for treatment. In this study, 13 medical fields and 204 symptoms, which are available on the website of many public hospitals associated with T.C. Ministry of Health and used to help patients choose the right medical treatment branch according to their symptoms, were examined using text mining and data science techniques. Based on the content of the text used, the closeness among the medical treatment areas was calculated and the words and symptoms confusing the patients the most while deciding the treatment area were determined. When analyzing the words, meaningless words were ignored, and a word cloud was created on the symptoms of the diseases. In order to calculate the closeness of medical fields, 13x186 binary data was created, indicating whether each symptom exists. Later, the medical fields on this data set were clustered according to the symptoms using agglomerative hierarchical clustering algorithms and the proximity of medical treatment fields was found. In the results, the words that challenge patients the most and the text-based affinities of medical fields are shared. Reorganizing content of the official document used on the hospital websites using the results obtained on this study will help to reduce the number of appointments received from the wrong medical branches.

Keywords: Hospital appointment, Text mining, Data science, Disease symptoms, Health.

* Sorumlu Yazar: hakurban@gmail.com

1. Giriş

İnsan vücudu kompleks bir yapıdadır. Vücudun belirli bir bölgesinde meydana gelen bir rahatsızlık, vücudun başka bir bölgesinde/bölgelerinde, o bölge ile doğrudan ilgisi olmayan bir sorunun/sorunların sonucu olarak ortaya çıkabilmektedir. Örneğin; mide bulantısı yaşayan bir hastanın hastalığı, dahiliye, intaniye, hariciye ve bevliye bölümleriyle ilgili bir problemi gösterebilmektedir. Ya da baş dönmesi belirtisinin; dahiliye, kulak burun boğaz (KBB), nöroloji veya göz bölümleri ile ilişkisi bulunabilmektedir. İnsan vücudundaki herhangi bir hastalık belirtisinin birçok tıbbi alanla ilgili olması, hastalarının tedavi için gitmesi gereken tıp alanını seçerken hata yapmalarına sebep olmaktadır. Ayrıca bu durum, hastane kaynaklarının optimal bir şekilde kullanılmasına engel olup, sağlık hizmetlerinin verilmesinde çeşitli problemlere yol açmaktadır [1]. İngiltere'de yapılan bir çalışmaya göre bir hastanın yanlış bir doktora gitmesinin masrafı yaklaşık olarak 75-100 İngiliz poundu olarak hesaplanmıştır [2]. Bir hastanın kendisinin rahatsızlığı ile alakalı olmayan bir tedaviyi beklemesi, o tedavi için bekleyen hastaların bekleme süresini arttırırken, bir yandan da tedavisinin daha masrafları bir hale gelmesine sebep olmaktadır. Tedavinin en yararlı olduğu zaman diliminin göz ardı edilmesinin ne vahim sonuçlar doğuracağını, Çin'de meydana gelen ve bütün dünyayı etkisi altına alan Korona (Covid'19) virüsünün sonuçlarına bakıldığı zaman daha iyi bir şekilde anlaşılabilmektedir.

Büyük veri ve yapay zekâ (YZ) tabanlı akıllı randevu sistemleri, günümüzde yaygın ve aktif olarak kullanılmamasına rağmen, akademik olarak son zamanlarda YZ tabanlı hastaları doğru tedavi alanlarına yönlendirebilecek sistemler üzerine çalışmalar yapılmaktadır [3-5]. YZ tabanlı bu tür sistemler hastaları en doğru branşa yönlendirebilmelerin yanında, hastanelerin kaynaklarının daha optimal olarak kullanılmasına katkıları olmaktadır. Kurulan sistem sayesinde, sağlık giderleri azaltılıp, sağlık çalışanlarının ve hastaların memnuniyetlerinin arttırılması sağlanmaktadır. Metin madenciliği [6]; sağlık alanı dahil olmak üzere birçok alanda metinlerde saklı ve önemli bilgilerin ortaya çıkarılması ve metinlerın geliştirilmesi için kullanılmaktadır [7-11]. Bu çalışmada T.C. Sağlık Bakanlığının, birçok kamu hastanesinin resmî sitesinde, hastalara randevu alırken yardımcı olmak amacıyla kullandığı, hastalık belirtileri ile tıbbi tedavi alanlarıyla ilgili metin, metin & veri madenciliği, makine öğrenimi ve veri bilimi teknikleriyle incelenmiştir. Hastanelerin bahsedilen dökümanı bu çalışmadan çıkan sonuçlara göre yenilemeleri sağlık sisteminin daha verimli bir şekilde çalışmasına faydası olacaktır. Bu çalışmada cevabı aranan bazı araştırma soruları aşağıdaki gibidir:

- En sık kullanılan ve birden fazla tedavi alanı ile ilgisi olan ve hastaları randevu alırken, alan seçiminde zorlayan kelimeler ve hastalık belirtileri nelerdir?
- Tıbbi tedavi alanlarının birbirine yakınlığını/benzerliğini, verilen metin veri setini kullanarak, makine öğrenimi algoritmaları ile anlayabilir miyiz?
- Belirtilere göre tıbbi tedavi alanları arasındaki korelasyon nasıl görülmektedir?

Bu çalışmada yazılan bilgisayar programları, verili bilimi ve hesaplama bilimleri alanını içeren birçok alanda her geçen gün yaygınlığını arttıran R programlama dili ile yazılmıştır. Çalışma için oluşturulan kodlar kamuya açık bir şekilde GitHub aracılığı ile paylaşılmıştır. Ayrıca, her bir tıbbi tedavi alanı için metinde verilen herbir belirtinin (186 belirti) var olup olmadığını gösteren dönem vektörleri (term vector) oluşturulduktan sonra, bu dönem vektörleri birleştirilip ilişkisel yeni bir veri seti oluşturulmuştur. Bu veri seti ikili (binary) seyrek veri seti olup, eğitimsel olarak ve araştırma amacıyla kullanılabilmesi amacıyla kamuya açık bir şekilde paylaşılmıştır [12].

2. Materyal ve Metot

2.1. Hiyerarşik Kümeleme

Bilgisayarda öğrenme (makina öğrenmesi); eğitimsiz öğrenme (unsupervised learning) [13] eğitimli öğrenme (supervised learning) [14], yarı eğitimli öğrenme (semi-supervised learning) [15] ve pekiştirmeli öğrenme (reinforcement learning) [16] olmak üzere dört ana başlık altında toplanmaktadır. Bu çalışmada metne bağlı olarak tıbbi alanlar arasındaki ilişkileri (yakınlık/uzaklık, benzerlik/ farklılık) ortaya çıkarıp, hastaları randevu alırken en çok zorlayan alan/alanların tespiti için, eğitimsiz öğrenmenin ana yöntemlerinden biri olarak bilinen kümeleme algoritma metotlarından faydalanılmıştır. Benzer yaklaşımla, farklı alanlarda yapılandırılmamış metin üzerinde döküman veri matrisi oluşturularak, bu matrise dayalı kelimelerin kümelenmesi yaygın bir şekilde uygulanmaktadır [17-20]. Kümeleme algoritmalarının performansları ve sonuçları farklı veri setleri üzerinde değişiklikler göstermektedir [21]. Örneğin; en popüler kümeleme algoritmalarından biri olan k-means kümeleme algoritması, küre şeklindeki veriler üzerinde daha iyi performans göstermesine rağmen, hiyerarşik kümeleme algoritmaları daha çok hiyerarşik yapıların olduğu veri setleri üzerinde daha iyi performans göstermektedir. Algoritmaların performanslarının değişiklik göstermesinin sebebi, tasarım şekillerinden kaynaklanmaktadır. Örneğin; k-means algoritması istenilen sayıda küme oluştururken, yinelemeli olarak (1) denkleminde verilen konveks fonksiyon olan kare toplamı hata fonksiyonunu (maliyet fonksiyonu) J'yi minimize etmektedir. (1) denkleminde: k: küme sayısını, x: veri setindeki herhangi bir veri noktasını, $C_j.v$: j-kümesinin merkez noktasını (centroid), $C_j.X$: j- kümesindeki elamanları göstermektedir.

$$J = \sum_{j}^{k} \sum_{x \in C_j.X} min||\boldsymbol{x} - C_j.v||_2^2 \qquad (1)$$

Verinin şekline ve türüne göre tasarlanmış birçok popüler kümeleme algoritması bulunmaktadır [22]. Elimizdeki verinin seyrek yapıda olması, boyutundan ve gösterim şeklinin dendrogram adı verilen, ağaç benzeri yapılarla kolay anlaşılabilir olması nedeniyle hiyerarşik kümeleme yöntemleri tercih edilmiştir. $\Delta = \{x_0, \dots, x_n\}$ elimizdeki veri seti ve her bir veri noktası $x_i \in \Re^d$, d-boyutlu bir vektör olsun. Hiyerarşik kümeleme algoritmaları, yinelemeli olarak Δ'yı $C_1, C_2, \dots, C_\ell$ **(** $l \leq n$**)** tane kümeye parçalamaktadır. Hiyerarşik kümeleme algoritmaları, diğer kümeleme algoritmalarından farklı olarak, kümeleme işlemine başlamadan önce, küme sayısı verilmeden,

aglomeratif ve ayırcı olarak adlandırılan iki farklı yaklaşımla veriyi kümelere ayırmaktadır. Ayırıcı hiyerarşik kümeleme yöntemleri, aglomeratif hiyerarşik kümeleme metotlarına göre daha az yaygınlıkla kullanılmakta olunup, yukarıdan aşağıya (top-down) hiyerarşik kümeleme yöntemi olarak da bilinmektedir. Ayrıca, ayırcı hiyerarşik kümelemede, aglomeratif kümelemeden farklı olarak ilk adımda bütün veri noktaları bir küme olarak kabul edilip, her bir kümede bir veri noktası kalana kadar Δ parçalanmaktadır. Hiyerarşik kümelemede istenilen sayıda küme elde edilmesi işlemi dendrogram'ın kesilmesi olarak adlandırılmaktadır.

2.1.1. Aglomeratif (Agglomerative) Hiyerarşik Kümeleme

Aşağıdan yukarıya hiyerarşik kümeleme algoritması olarak da bilinmektedir. Aşağıda verilen algorithm 1 aglomeratif hiyerarşik küme algoritmasının detaylarını göstermektedir. Algoritmanın girdi parametreleri veri matrisi Δ, yakınlık/uzaklık/benzerlik/farklılık metriği M (satır 1) ve çıktısı n veri noktasından bir küme oluşumunu gösteren dendrogram'dır (satır 2). İlk adımda (satır 4) her bir veri noktası bir küme olarak kabul edilip, veri noktaları arasındaki uzaklığı/yakınlığı/benzeri/farklığı gösteren nxn uzaklık/yakınlık (benzerlik/farklılık) matrisi T hesaplanır. Daha sonra yinelemeli olarak T baz alınarak her bir adımda küme sayısı bir azaltılarak sadece bir küme elde edilene kadar en yakın iki küme birleştirilip, T güncellenir (satır 6-11). Aglomeratif hiyerarşik kümelemenin sonucunu kullanılan yakınlık/uzaklık/benzerlik/farklılık metriği M ve herbir adımda uzaklık/yakınlık (benzerlik/farklılık) matrisi T*'nin* güncellenme şekli belirlemektedir. Bu yöntemler linkage yöntemeleri olarak adlandırılmakta olunup, kullanılan verinin yapısına göre avantajlı ve dezavantajlı tarafları bulunmaktadır. Bu yöntemlerden en yaygın olanlarından bazıları detaylı olarak aşağıda açıklanmıştır. Bunların dışında Average, Ward, Mean, Mcquitty, Median diğer bilinen linkage metotlarıdır.

Algorithm 1: Δ üzerinde Aglomeratif Hiyerarşik Kümeleme

1. **INPUT** veri seti Δ, uzaklık/yakınlık (benzerlik/farklılık) metriği M, örneğin; $d: \Delta^2 \rightarrow \Re_{\geq 0}$
2. **OUTPUT** n-kümeden bir küme oluşumunu gösteren dendrogram
3. %% d × d uzaklık/yakınlık (benzerlik/farklılık) matrisi T'yi %% hesapla
4. $\Delta_{prox} \leftarrow \Delta$
5. %% Her bir veri noktası bir küme olsun.
6. **repeat**
7. En yakın/benzer iki kümeyi birleştir.
8. %% Δ_{prox} linkage tekniklerine göre güncellenir.
9. %% Örneğin; single, complete, ward.
10. Δ_{prox}'u güncelle
11. **until** Sadece bir tane küme kalana kadar.

2.1.1.1. Single (Min)

En yakın iki küme birleştirildikten sonra, bu iki birleştirilen kümelerin diğer kümelere olan uzaklığı, bu iki birleştirilen kümeden yakın/benzer olanına göre hesaplanıp, uzaklık/yakınlık (benzerlik/farklılık) matrisi T güncellenir. Örneğin; C_i ve C_j iterasyon (t)-de birleştirilen iki küme ve C_p iterasyon (t)'de bu iki küme dışında başka bir küme olsun. T güncellenirken C_i ve C_j'nin C_p'ye olan uzaklığı single linkage ile şu şekilde hesaplanmaktadır.

$$min_{i,j,p\leq|\Delta|}\left[d(C_i, C_p), d(C_j, C_p)\right] \quad (2)$$

2.1.1.2. Complete (Max)

En yakın iki küme birleştirildikten sonra, bu iki birleştirilen kümelenin, diğer kümelere olan uzaklığı, bu iki birleştirilen kümeden uzak/farklı olanına göre hesaplanıp, uzaklık/yakınlık (benzerlik/farklılık) matrisi T güncellenir. Örneğin; C_i ve C_j iterasyon (t)-de birleştirilen iki küme ve C_p iterasyon (t)'de bu iki küme dışında başka bir küme olsun. T güncellenirken C_i ve C_j'nin C_p'ye olan uzaklığı complete linkage ile aşağıdaki gibi hesaplanmaktadır.

$$max_{i,j,p\leq|\Delta|}\left[d(C_i, C_p), d(C_j, C_p)\right] \quad (3)$$

2.1.1.3. Centroid (Mean)

En yakın iki küme birleştirildikten sonra, bu iki birleştirilen kümelenin diğer kümelere olan uzaklığı, bu iki birleştirilen kümenin merkezlerinin, bu iki kümenin dışındaki kümelere olan uzaklıkları baz alınarak, uzaklık/yakınlık (benzerlik/farklılık) matrisi T güncellenir (yakın olanına göre). Örneğin; $C_i.v$ ve $C_j.v$ iterasyon (t)-de birleştirilen iki kümenin merkezi ve $C_p.v$ iteresyon t'de bu iki küme dışında bir küme olsun. T güncellenirken C_i ve C_j'in C_p'ye olan uzaklığı centroid linkage ile şu şekilde hesaplanmaktadır.

$$min_{i,j,p\leq|\Delta|}\left[d(C_i.v, C_p.v), d(C_j.v, C_p.v)\right] \quad (4)$$

2.2. Metin Madenciliği

Metin madenciliği (text mining); yapılandırılmamış metin içerisindeki saklı, önemli ve anlamlı olan kalıpları ortayaya çıkarabilmek için, yapılandırılmamış metnin yapılandırılmış biçime dönüşütürülme sürecidir [23,24]. Metin madenciliği teknikleriden; organizasyonel araştırma [25], finans [26], turizm (turistik mekanlar için yapılan incelemelerin anlaşılması) [27], sosyal medya [28] gibi yapılandırılmamış metnin bulunduğu birçok alanda faydalanılmaktadır. [29], metin madenciliği tekniklerinin literatünü içermektedir. Metin madenciliğinin kullanım alalarından bir tanesi, yapılandırılmamış bir metinde kullanılan en önemli kelimelerin tespit edilmesidir [30]. Tag cloud ya da world cloud [31] olarak bilinen metin bulutu, bir metindeki en önemli kelimelerin tespiti için edat ve bağlaç benzeri anlamsız kelimelerin veriden çıkarıldıktan sonra bu metinde bulunan kelimelerin görsel olarak çizdirilmesidir. Bu tür görsellerde kelimeler, kullanım sıklıklarına göre boyutlandırılmaktadır. Hastalık belirtileri metni içerisinde, en sık kullanılan kelimelerin kelime bulutu, hastaların tedavi alanını seçerken en çok zorlanabileceği kelimelerin tespitinin yapılabilmesi için kullanılmıştır.

2.3. Veri Seti

Bu çalışmada T.C. Sağlık Bakanlığına bağlı hastanelerin, resmî internet sitelerinde, hastalara tedavi için randevu alırken doğru tıbbi tedavi alanı seçerken yardımcı olmak için kullandıkları *"Nerde muayene olmalıyım?"* adlı metin, veri seti olarak kullanılmıştır. Bu veri seti, 13 tane tıbbi tedavi alanı hakkında 204 tane belirtiyi metin olarak içermektedir. Bu veri setindeki tıbbi alan isimleri ve her bir alanla ilgili verilen belirti

sayıları Tablo 1'de gösterilmiştir. Kullanılan veri setinin benzer versiyonları birçok kamu hastanesinin sitesinde bulabilmekle birlikte, bu çalışmada kullanılan veri seti T.C. İstanbul İl Sağlık Müdürlüğü Sancaktepe Şehit Prof. Dr. İlhan Varank Eğitim ve Araştırma Hastanesinin web sitesinden alınmıştır [32].

Bu çalışmada, tıbbi tedavi alanların birbirine yakınlığını/benzerliğini hesaplayabilmek için, her bir belirtinin her bir alanda var olup olmadığını gösteren ilişkisel veri seti oluşturulmuştur. Bu veri seti oluşturulurken metin içinde verilen 204 belirtiden farklı tıbbi tedavi alanlarında kullanılan aynı belirtiler sadece bir kez sayılarak toplam belirti sayısı 186'ye indirilip, $13x186$ döküman matrisi elde edilmiştir (1: Belirtinin olduğunu, 0: Belirtinin olmadığını göstermektedir). Oluşturulan bu veri seti seyrek (sparse) veri setidir. Tablo 2'de bu veri setinin küçük bir örneği ve Şekil 1'de belirti sayılarının alanlar üzerindeki dağılımı gösterilmiştir.

Tablo 1: Veri Seti Özeti: Her Bir Tıbbi Alan İçin Verilen Belirti Sayıları.

Tıbbi Tedavi Alanı	Belirti Sayısı
Dahiliye (İç Hastalıkları)	10
Kulak Burun Boğaz (KBB)	20
Enfeksiyon Hastalıkları	14
Genel Cerrahi (Hariciye)	23
Nöroloji	17
Dermatoloji (Cildiye)	21
Ortopedi	13
Üroloji	18
Psikiyatri	24
Göz Hastalıkları	11
Kardiyoloji	5
Plastik Cerrahi	19
Göğüs Hastalıkları	9

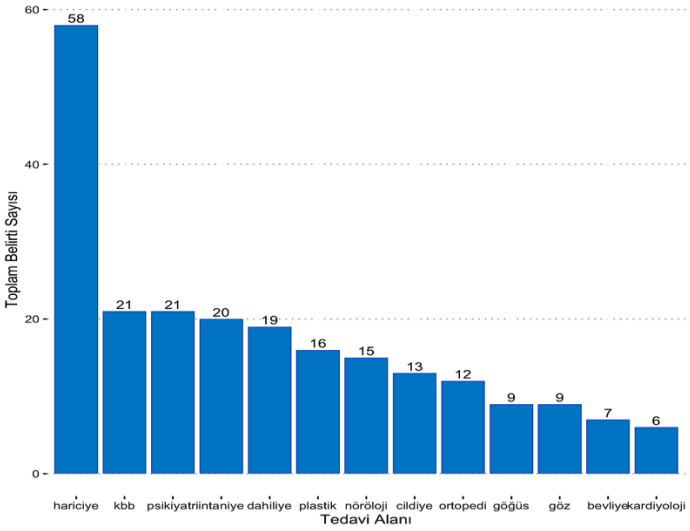
Tablo 2: Veri Setinin Küçük Bir Örneği.

	ağız.kuruluğu	göğüs.ağrısı	kol.ağrısı	sık.idara.çıkma	karın.ağrısı	mide.ağrısı	mide.eksimesi	gastrit
dahiliye	1	1	1	1	1	1	1	1
kbb	0	0	0	0	0	0	0	0
intaniye	0	0	0	0	1	0	0	0
hariciye	0	0	0	0	1	1	0	0
nöroloji	0	0	0	0	0	0	0	0
cildiye	0	0	0	0	0	0	0	0
ortopedi	0	0	0	0	0	0	0	0
bevliye	0	0	0	1	0	0	0	0
psikiyatri	0	0	0	0	0	0	0	0
göz	0	0	0	0	0	0	0	0
kardiyoloji	0	0	0	0	0	0	0	0
plastik	0	0	0	0	0	0	0	0
göğüs	0	1	0	0	0	0	0	0

3. Araştırma Sonuçları ve Tartışma

3.1. Kelime Bulutu ile Metnin İncelenmesi

Kullanılan metin içerisinde hangi kelimelerin/semptomların hastaları randevu alırken zorladığını veya yanlış alanlardan randevu almalarına sebep olduğunu tespit etmek için metin içerisinde geçen kelimelerin kullanılma sıklığı kelime bulutu (word cloud) görseli çizdirilerek gözlemlenip, analiz edilmiştir. Bu görsel çizdirilmeden önce metin içeresinde geçen edat bağlaç gibi anlamsız kelimeler (stop words) öncelikle tespit edilip bu sayımın dışında bırakılmıştır. Bu kelimelerin listesi Tablo 3'te verilmiştir. Anlamsız kelimelerin sayım dışı tutulması dışında, aynı anlama sahip farklı ekler almış kelimeler aynı kelime olarak sayılmıştır. Örneğin; "bozuklukları", "bozukluğu" kelimeleri "bozukluk" kelimesi olarak sayılmıştır. Şekil 2'de metin içerisinde en çok kullanılan 10 kelimenin kullanılma sıklıkları ve metin içerisinde iki defadan fazla kullanılan kelimelerin kelime bulutu gösterilmiştir. Anlamsız kelimelerinin silinmesinden sonra geriye 410 tane kelime kalmıştır. Bu kelimelerin 307 tanesi 1 defa, 49 tanesi 2 defa, 26 tanesi 3 defa ve geriye kalan 28 kelime 4 ve üzeri sayıda metin dosyasında karşımıza çıkmaktadır. "ağrı" ve "bozukluk" kelimeleri açık ara diğer kelimelerden daha fazla sayıda 38, 21 defa metin içerisinde kullanılmıştır. En çok kullanılan 10 kelimenin alanlarla ve belirtiler içerisinde kullanılma şekli ile ilgili detaylı analiz aşağıda verilmiştir.



Şekil 1: Oluşturulan Dönem Matrisinde Belirti Sayılarının Alanlara Göre Dağılımı.

Tablo 3: Metinden Çıkarılan Kelimeler.

Metinden Silinen Kelimeler (Stop Words)
"bir", "eden","herhangi", "iki", "hastalıklarda", "ile","gibi", "gibi", "kendini","olarak","hastalıkları","hastalığı", "hastalık","şikayetleri", "fakat", "için", "ya da","olarak", "buna", "veya"

Tablo 5: Bozukluk Kelimesinin İstatistikleri ve Analizi.

Bozukluk Tipi	Tıbbi Tedavi Alanı
Burunda Şekil Bozukluğu	KBB, Plastik Cerrahi
Konsantrasyon	Genel Cerrahi
Tırnak	Cildiye
Sosyalleşme, Davranış, Uyku, Kaygı, Kişilik, Bipolar, Şizoeffektif, Karşı gelme	Psikiyatri
Görme netliği bozukluğu	Göz Hastalıkları
Kulak şekil / bozuklukları	Plastik Cerrahi

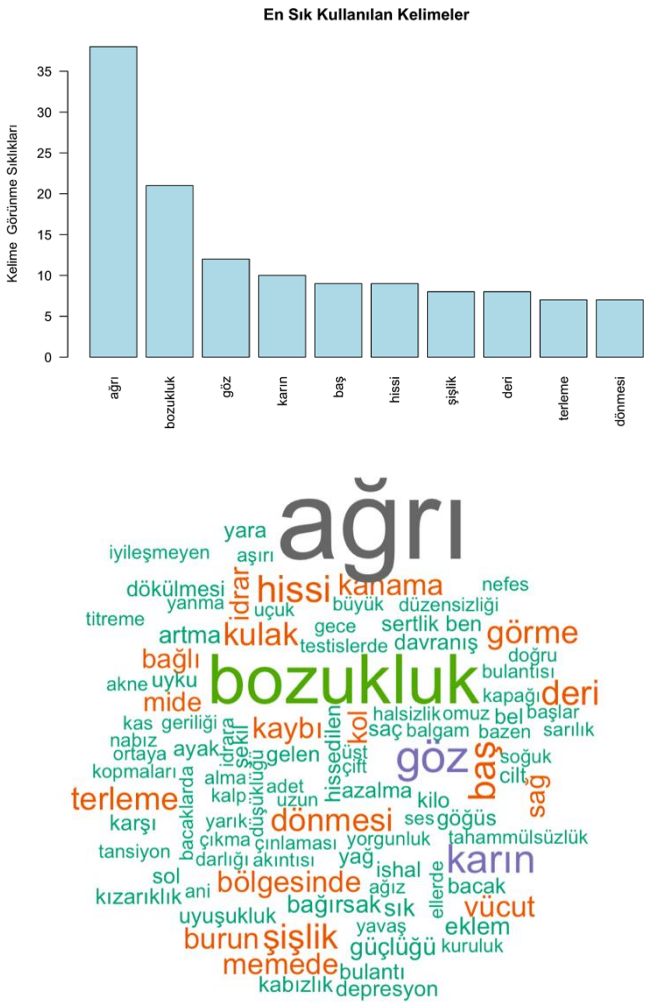
Tablo 4: Ağrı Kelimesinin İstatistikleri ve Analizi.

Bölge/Uzuv	Tıbbi Tedavi Alanı
Göğüs	Dahiliye, Göğüs Hastalıkları
Karın	Dahiliye, Enfeksiyon Hastalıkları, Genel Cerrahi
Mide	Dahiliye
Baş	Dahiliye, Nöroloji, Göz Hastalıkları
Boğaz	KBB
Kulak	KBB
Ayak	KBB
Meme	Genel Cerrahi
Sırt	Genel Cerrahi
Göbek	Genel Cerrahi
Kasık	Genel Cerrahi
Yüz	Nöroloji
Eklem	Ortopedi
Bacak/Boyun/Fıtık/Kalça	Ortopedi
Kas	Ortopedi
Bel	Ortopedi
Testis	Üroloji
Göz	Göz Hastalıkları
Kol	Kardiyoloji

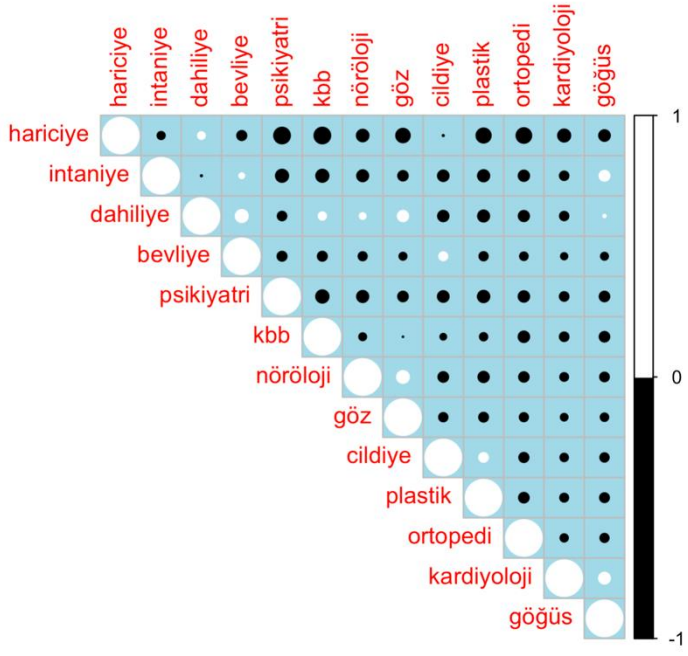
Ağrı: Ağrı kelimesi bölgesel ya da uzuvsal 19 farklı ağrıyı anlatmak için kullanılmış olup, bu belirtiler 10 farklı tıbbi tedavi alanı ile ilişkilendirilmiştir. Genel Cerrahi için 5 farklı belirtiyi; Dahiliye ve Ortopedi bölümleri için 4 farklı belirtiyi; KBB için 3; Nöroloji ve Göz Hastalıkları için 2; Göğüs Hastalıkları, Enfeksiyon Hastalıkları**,** Üroloji ve Kardiyoloji Bölümleri içinde 1'er belirtiyi anlatmak için kullanılmıştır. Bu belirtiler içerisinde karın ağrısı ve baş ağrısı 3'er tedavi branşı ile ilişkilendirilmesi sebebiyle seçim konusunda hastaları en çok zorlayacak belirtiler olarak göze çarpmaktadır. Ayrıca göğüs ağrısı iki alanla ilişkilendirilmiştir. Tablo 4'te ağrı kelimesinin bölümlere göre detaylı analizi gösterilmiştir.

Bozukluk: Bozukluk kelimesi en çok kullanılan ikinci kelime olmasına rağmen hastaları tedavi alanı seçimi yaparken zorlayacak kelimeler arasında yer almamaktadır. 13 bozukluk belirtisin 8 tanesi sadece Psikiyatri bölümüyle ilgilidir. Plastik Cerrahi ile ilgili 2; Cildiye, Genel Cerrahi, Göz Hastalıkları ve KBB ile ilgilide 1'er belirti bulunmaktadır. Sadece "*burunda şekil bozukluğu"* tedavisi ile birden fazda alan ilgilenmektedir. Tablo 5'te "*bozukluk"* kelimesinin daha ayrıntılı olarak istatistikleri verilmiştir.

Göz, Karın, Baş, Hissi, Şişlik, Deri, Terleme, Dönmesi, Kelimelerinin Analizi: En çok kullanılan kelimeler analizinin sadeliği ve daha anlaşılır olması için, ağrı ve bozukluk kelimelerinden sonra en çok kullanılan diğer kelimelerin incelenmesi tablo halinde verilmiştir. Bu bölümde bu kelimeler ile ilgili önemli olan noktalara değinilmiştir. Tablo 6, *ağr*ı ve *bozukluk* kelimelerinden sonra en sık kullanılan 8 kelimenin analizini içermektedir.



Şekil 2: En Sık Kullanılan Kelimeler ve Kelime Bulutu.



Şekil 3: Tedavi Alanları Arasındaki Kendall Korelasyou.

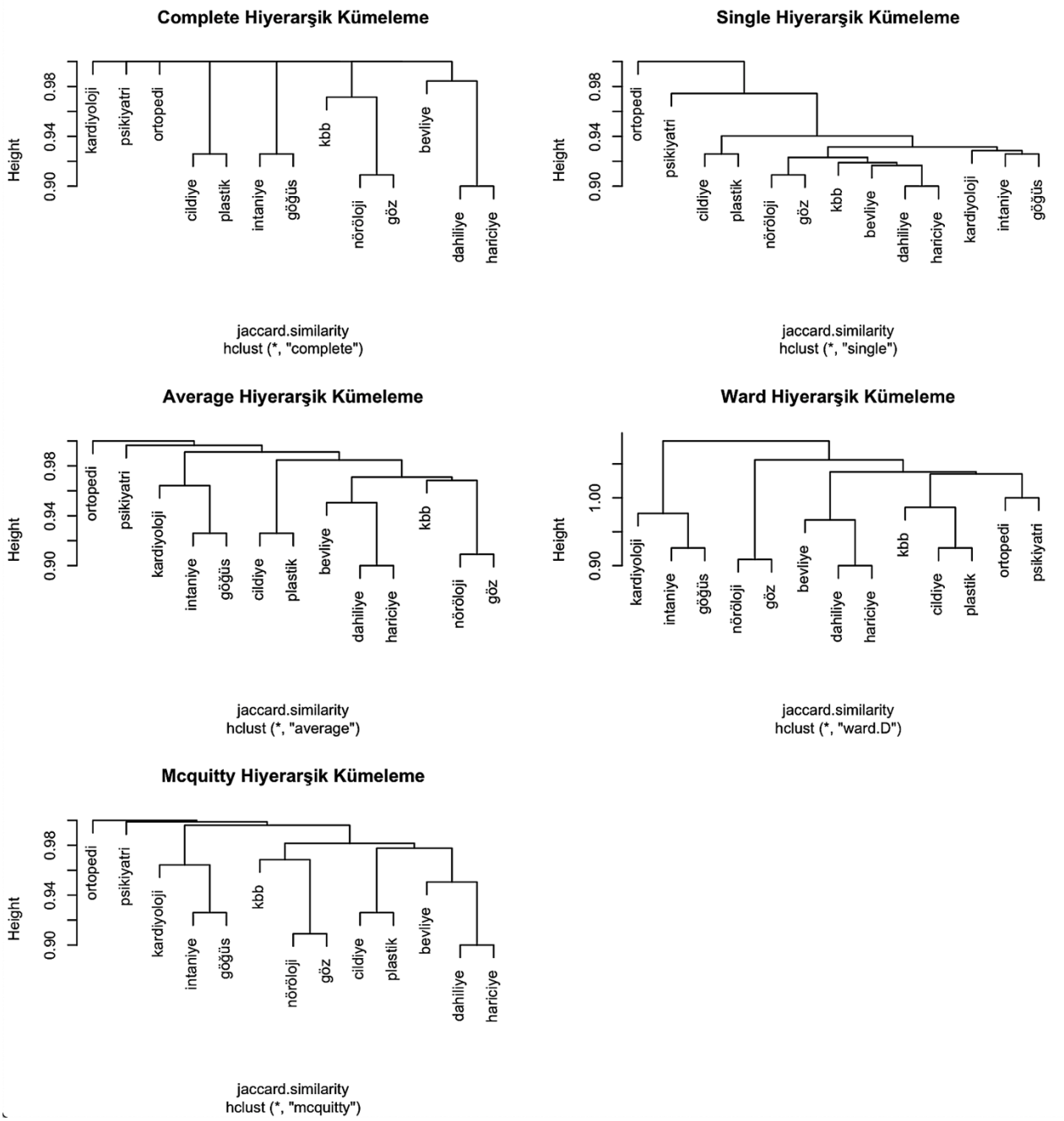
3.2. Metne Dayalı Tedavi Alanlarının Birbirine Yakınlıklarının Ölçülmesi

Metinde bulunan ve hastaları randevu alırken zorlayan belirtilerin ve kelimelerin tespiti ve analizini yaptıktan sonra, verilen metne göre, hastalık belirtilerinin her bir tedavi alanında var olup olmadığını gösteren döküman matrisi oluşturulmuştur. Öncelikle, bu ikili veri seti üzerinde Kendall korelasyon katsayısı hesaplanarak; tıbbi alan metinlerinin birbiriyle korelasyonu gözlemlenmiştir. Şekil 3'te görüldüğü üzere metin bazlı alanlar arasında aşırı olmayan korelasyonlar gözlenmektedir. Bunun yanında, tıbbi alanlar bu doküman matrisi kullanılarak, hiyerarşik kümeleme algoritmaları üzerinde kümelere ayrılıp, metin bazlı alanların birbirine yakınlığı ölçümü yapılmıştır. Buradaki amacımız birbirine yakın kümeleri tespit etmektir. Her bir küme bir tedavi alanını temsil ettiği için iki kümenin birbirine yakın olması, alanlar arasında metin benzerliğinin yüksek olmasını göstermektedir. Hiyerarşik kümeleme algoritmasının girdi parametrelerinden biri olan uzaklık/yakınlık matrisi olarak Jaccard yakınlık metriği kullanılmıştır. Jaccard metriğine göre alanlar arası metin bazlı benzerlik genel olarak düşük olarak görülmesine rağmen, birbirine en çok benzeyen alan metinlerinin tespiti için hiyerarşik kümeleme teknikleri kullanılmıştır. Şekil 4 oluşturduğumuz dönem matrisinin (term matrix) farklı aglomeratif hiyerarşik kümeleme yöntemleri ile kümeleme sonuçlarını göstermektedir. Linkage metotları olarak single, complete, average, ward and mcquitty kullanılmıştır. Şekil 4 incelendiğinde Ortopedi ve Psikiyatri için kullanılan metinlerin bütün sonuçlarda diğer alanlarla benzerliğinin düşük olduğu görülmektedir. Bu sonucun dışında metin bazlı alanlar arasında aşağıdaki belirtilen sonuçlar dikkat çekmektedir:

- İntaniye, Göğüs Hastalıkları ve Kardiyoloji bölümleri arasında metin bazlı benzerlik olmasına rağmen; Göğüs Hastalıkları, İntaniye branşları arasındaki metne bağlı benzerlik oranı Kardiyolojiye göre daha yüksektir.
- Nerdeyse bütün sonuçlarda Göz Hastalıkları ve Nöroloji bölümleri arasındaki belirti bazlı yakınlık diğer bölümlerin bu iki bölüme yakınlığından daha fazladır.
- Cildiye ve Plastik Cerrahi alanları bütün sonuçlarda birbirine en yakın kümeler olarak göze çarpmaktadır.
- Dahiliye ve Hariciyenin birbirine en yakın kümeler olduğu ve bu iki kümenin birleşimin en yakın kümesinin her zaman bevliye olduğu görülmektedir.
- KBB bölümüyle alakalı metnin en karışık metin olduğu görülmektedir. KBB kümesi bazen Nöroloji ve Göz Bölümlerinin birleşimin en yakın kümesi, bazen cildiye ve plastiğin birleşime en yakın küme, bazen de hariciye dahiliye ve bevliye kümelerinin birleşimine en yakın küme durumundadır.

Tablo 6: Ağrı ve Bozukluk Kelimeleri ile En Çok Kullanılan Kelimelerin İstatistikleri ve Analizi.

Kelime	Analiz
Göz	Göz kelimesi 4 tedavi alanını içeren 13 belirtti için kullanılmıştır. Bu belirtilerin 7 tanesi Göz Hastalıkları, 3 tanesi Genel Cerrahi, 2 tanesi Plastik Cerrahi ve 1 tanesi İntaniye branşlarıyla ilgilidir. Hiçbir belirti birden fazla branşla ilgili değildir.
Karın	3 belirti 3 alan ile ilişkilendirilmiştir. 2 belirti sadece Genel Cerrahi ile ilgi olmasına rağmen, karın ağrısı belirtisi Dahiliye, Enfeksiyon ve Genel Cerrahi Bölümleri ile ilişkilendirilmektedir.
Baş	2 belirti 4 branşla eşleşmektedir. Sorunun anlaşılmasını güçlendiren bir kelime olarak dikkat çekmektedir. Baş ağrısı Nöroloji, Dahiliye, ya da Göz Hastalıkları ile ilgili olabilmesine rağmen; Baş dönmesi, Dahiliye, KBB Nöroloji veya Göz Hastalıkları bölümleri ile alakası bulunabilmektedir.
Hissi	6 branşı içeren 12 belirti için kullanılmasına rağmen hiçbir belirtinin birden fazla alanla alakası olmadığı için, hastalara zorluk çıkaran kelimelerden bir tanesi değildir. His ile ilgili belirtiler; Genel Cerrahi, Nöroloji, Ortopedi, Üroloji, Göz Hastalıkları, Kardiyoloji ilgili bölümlerden birine işaret etmektedir.
Şişlik	8 belirti, 3 alan (Genel, Cerrahi Enfeksiyon, Üroloji) ile ilişkilendirilmiştir. Bu belirtilerden sadece bacaklardaki şişlik birden alanla ilgili bir sorunu göstermektedir (Genel Cerrahi, Enfeksiyon).
Deri	8 belirti, 3 branşla ilişkilendirilmiştir. Sadece deri kanseri belirtisi birden fazla tedavi alanı ile ilişkilendirilmiştir (Cildiye, Plastik Cerrahi).
Terleme	4 farkı terleme başlığı altında, 6 branşı içeren sorunlu kelimelerden bir tanesidir. Terleme belirti olarak Dahiliye, Enfeksiyon Hastalıkları, Genel Cerrahi, Cildiye, Göğüs Hastalıkları ile ilgili sorunlara işaret etmektedir.
Dönmesi	Kıl dönmesi ve baş dönmesi olmak üzere 2 belirti için kullanılmıştır. Baş dönmesi analizi baş kelimesi altında yapılmıştır. Kıl dönmesi sadece Cildiye branşı ile ilgili bir soruna işaret etmektedir.



Şekil 4: Tedavi Alanlarının Hastalık Belirtilerine göre Farklı Hiyerarşik Kümeleme Algoritmaları ile Kümelenmesi.

4. Sonuç

Bu çalışmada, T.C. Sağlık Bakanlığına bağlı birçok kamu hastanesinin, hastaların doğru tedavi alanından randevu alabilmelerine yardımcı olmak için kullandığı "Nerde muayene olmalıyım?" başlıklı metin, metin madenciliği, veri bilimi ve makine öğrenimi tekniklerini kullanarak analiz edilip, önemli görülen sonuçlar paylaşılmıştır. Öncelikle, metin veri seti kullanılarak, hastaları randevu alırken en çok zorlayan kelimeler (anahtar kelimeler, hastalık belirtikeri) kelime bulutu ile tespit edilip, gösterilmiştir. Daha sonra farklı hiyerarşik kümeleme algoritmları ve kendall korelasyonu kullanılarak, metin bazlı tıbbi tedavi alanlarının birbirine yakınlığı/benzerliği tespit edilip, tıbbi tedavi alanları için kullanılan açıklamaların hastaları doğru yönlendirebilmesi için yeterli olup olmadığı hesaplanmıştır. Kelime bulutu üzerinde hastaları en çok zorlayan kelimelerin ağrı, karın, baş, dönmesi, terleme ve şişlik kelimeleri olduğu görülmektedir. Bu kelimelerden oluşan baş dönmesi, baş ağrısı, karın ağrısı, bacak şişliği ve terleme belirtilerinin birçok tedavi alanıyla ilişkisi olmasından dolayı, hastaları randevu alırken en çok zorlayan belirtiler olduğu göze çarpmaktadır. Bu belirtilerin başka belirtilerle desteklenmesi, yanlış tedavi alanlarından alınan randevu sayısının düşürülmesine faydası olacaktır. Tıbbi tedavi alanları, belirti bazlı oluşturduğumuz döküman matrisi veri seti üzerinde aglomeratif hiyerarşik kümeleme yöntemleri ile guruplara ayrılıp, tedavi alanları için kullanılan metinlerin birbirine benzerliği ölçülmüştür. Daha iyi değerlendirme yapılabilinmesi için ayrıca Kendall korelasyonundan faydalanılmıştır. Oluşturulan dendrogramlar baz alındığında, Psikiyatri ve Ortopedi alanları için kullanılan metinlerin diğer

alanlar için kullanılan metinlere göre daha açıklayıcı olduğu görülmektedir. Çünkü, kümeler (tedavi alanları) birbirine uzaklıklarına göre, dendrogram üzerinde aşağıdan yukarıya doğru sıralanmışlardır. Bunun yanında KBB için kullanılan metnin farkı hiyerarşik kümeleme algoritmaları için farklı sonuçlar göstermesi sebebiyle en problematik açıklamaya sahip olduğu anlaşılmaktadır. Bunların yanında (Göz, Nöroloji), (İntaniye, Göğüs Hastalıkları), (Cildiye, Plastik Cerrahi), (Dahiliye, Hariciye) ikilileri birbirine en çok benzeyen metinlere sahiptirler. Bunlara ek olarak, Kardiyolojinin en çok benzerlik gösterdiği metin Dahiliye ve Hariciye metinleridir.

Bu çalışma baz alınarak "Nerde muayene olmayım?" metninin sağlık uzmanları tarafından tekrar düzenlenmesinin, yanlış randevu sayısının azaltılmasına faydaları olacaktır. Ayrıca, yapay zekâ tabanlı oluşturulacak randevu sistemlerinde, sistemlerin performansını artırabilecek sonuçlara ulaşılmıştır. Örneğin; hastaların tedavi alanı sistem tarafından belirlenirken, sorulacak sorular içerisinde hangi belirtilerin olmaması gerektiğinin tesbiti yapılmıştır. Son olarak, bu çalışma için üretilen seyrek (sparse) veri seti ve R programlama dilinde yazılan programlar eğitim ve araştırma amaçlı kullanılabilmesi için kamuya açık bir şekilde (open source) paylaşılmıştır. Bu çalışmanın devamında buradaki bilgiler kapsamında yapay zekâ tabanlı hastane randevu sisteminin kurulması hedeflenmektedir.

Kaynakça

[1] Yavuz, İ., & Cagiltay, N. E. (2019). E-nabız Mobil Sağlık Uygulamasına Yönelik Kullanıcı Değerlendirmesi. *Hacettepe Sağlık İdaresi Dergisi*, *22*(2), 375-388.

[2] Vos, T., Barber, R. M., Bell, B., Bertozzi-Villa, A., Biryukov, S., Bolliger, I., ... & Duan, L. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, *386*(9995), 743-800.

[3] Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105-120.

[4] Zang, Y., Zhang, F., Di, C. A., & Zhu, D. (2015). Advances of flexible pressure sensors toward artificial intelligence and health care applications. *Materials Horizons*, *2*(2), 140-156.

[5] Yu, K. H., & Andrew, L. (2018). Beam, and Isaac S. Kohane. *Artificial intelligence in healthcare. Nature biomedical engineering*, *2*(10), 719-731.

[6] Tan, A. H. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases* (Vol. 8, pp. 65-70). sn.

[7] Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

[8] Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. Springer Science & Business Media.

[9] Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with R* (Vol. 65). New York, NY: Springer.

[10] Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., & Paris, C. (2015). Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)*, *47*(4), 1-39.

[11] Mittermayer, M. A. (2004). Forecasting intraday stock price trends with text mining techniques. In *37th Annual Hawaii International Conference on System Sciences. Proceedings of the (pp. 10-pp). IEEE*

[12] https://github.com/hasankurban/Hospital-Appointment

[13] Kurban, H., Jenne, M., & Dalkilic, M. M. (2017). Using data to build a better EM: EM* for big data. *International Journal of Data Science and Analytics*, *4*(2), 83-97.

[14] Mohsen, H., Kurban, H., Zimmer, K., Jenne, M., & Dalkilic, M. M. (2015). Red-rf: Reduced random forest for big data using priority voting & dynamic data reduction. In *2015 IEEE International Congress on Big Data* (pp. 118-125). IEEE.

[15] Uylaş S., N. (2018). A collective learning approach for semi-supervised data classification. *Pamukkale University Journal of Engineering Sciences*, *24*(5).

[16] Buşoniu, L., Babuška, R., & De Schutter, B. (2010). Multi-agent reinforcement learning: An overview. In *Innovations in multi-agent systems and applications-1* (pp. 183-221). Springer, Berlin, Heidelberg.

[17] Beil, F., Ester, M., & Xu, X. (2002). Frequent term-based text clustering. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 436-442).

[18] Liu, T., Liu, S., Chen, Z., & Ma, W. Y. (2003). An evaluation on feature selection for text clustering. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 488-495).

[19] Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In *Mining text data* (pp. 77-128). Springer, Boston, MA.

[20] Suarez-Tangil, G., Tapiador, J. E., Peris-Lopez, P., & Blasco, J. (2014). Dendroid: A text mining approach to analyzing and classifying code structures in android malware families. *Expert Systems with Applications*, *41*(4), 1104-1117.

[21] Jenne, M., Boberg, O., Kurban, H., & Dalkilic, M. (2014). Studying the milky way galaxy using paraheap-k. *Computer*, *47*(9), 26-33

[22] Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), 645-678.

[23] Jo, T. (2019). Text mining. *Studies in Big Data. Cham: Springer International Publishing*.

[24] Tan, A. H. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases* (Vol. 8, pp. 65-70). sn.

[25] Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational research methods*, *21*(3), 733-765.

[26] Pejić Bach, M., Krstić, Ž., Seljan, S., & Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. *Sustainability*, *11*(5), 1277.

[27] Aksu, M. Ç., & Karaman, E. (2020). FastText ve Kelime Çantası Kelime Temsil Yöntemlerinin Turistik Mekanlar İçin Yapılan Türkçe İncelemeler Kullanılarak Karşılaştırılması. *Avrupa Bilim ve Teknoloji Dergisi*, (20), 311-320.

[28] Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J*, *2*(1), 127-133.

[29] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of

text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.
[30] Koyun, A., & Yangeç, D. (2018). Veri Madenciliği Teknikleri Yardımıyla Otel Yorumlarından Anahtar Kelimeler Keşfi. *Avrupa Bilim ve Teknoloji Dergisi*, (14), 261-268.
[31] Dunaiski, M., Greene, G. J., & Fischer, B. (2017). Exploratory search of academic publication and citation data using interactive tag cloud visualizations. *Scientometrics*, *110*(3), 1539-1571.
[32] https://sancaktepeah.saglik.gov.tr/TR,250181/nereye-muayene-olmaliyim.html.