ESTUDAM Bilişim Dergisi
Cilt 4, Sayı 1, 24-28, 2023

Journal of ESTUDAM Information
Volume 4, Issue 1, 24-28, 2023

DOI: 10.53608/estudambilisim.1224150

(Geliş Tarihi / Received Date: 02.01.2023, Kabul Tarihi/ Accepted Date: 27.01.2023)

Araştırma Makalesi

Pubmed Platformunda Cerrahi Alanında Yayınlanmış Makalelerin Metin Madenciliği Teknikleri ile İncelenmesi

Seher KIZILTEPE[*1], Eyyüp GÜLBANDILAR[1], Faik YAYLAK[2]

[1]Kütahya Bilim ve Sanat Merkezi, Kütahya,
Orcid ID: https://orcid.org/0000-0001-6456-3484
[2]Eskişehir Osmangazi Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Eskişehir,
Orcid ID: https://orcid.org/0000-0001-5559-5281
[3]Kütahya Sağlık Bilimleri Üniversitesi, Tıp Fakültesi, Cerrahi Tıp Bilimleri Bölümü, Genel Cerrahi A.B.D. Kütahya,
Orcid ID: https://orcid.org/ 0000-0002-1216-0429

Anahtar Kelimeler:
Metin Madenciliği,
Metin Sınıflandırma,
Metin Analizi,
Doğal Dil İşleme,
Dijital Veri

Özet: Dijital çağ olarak nitelendirilen bu çağda, iletişim teknolojilerinin sunduğu hizmetler ile dijital verilerin hem önemi hem de sayısı her geçen gün hızla artmaktadır. Karmaşık yapıdaki metinlerden anlamlı kelimeleri çıkarmak ve bilgiye ulaşmak için kullanılan en yaygın yöntemlerden birisi de Metin Madenciliği (MM) yöntemleridir. MM çalışmaları birçok alanda olduğu gibi tıp alanında da yaygın kullanılmaktadır. Bu çalışmanın amacı, İngilizce dilindeki bir tıp veri tabanı olan Pubmed platformu üzerinde bulunan ve cerrahi alan ile ilgili yayınlanmış makalelerden MM yöntemleri kullanılarak cerrahi alanındaki çalışmaların yönelimi hakkında fikir sahibi olmaktır. Aynı zamanda bu alanda yapılmış çalışmaların özetleri üzerinde MM kullanılarak anahtar kelimeler elde etmek ve bu kelimelerin frekans değerlerini görsel olarak sergilemektir. Çalışmanın veri setini oluşturan text dosyası üzerinde önce metin ön işleme daha sonra da metin analiz yöntemleri kullanılarak metin içerisinden yaygın olarak kullanılan beş adet anahtar kelime üretilmiştir. Üretilen anahtar kelimelerin frekans değerleri görselleştirilerek grafik ve kelime bulutu başarılı bir şeklinde ortaya konulmuştur.

Research Article

Examination of Articles Published in the Field of Surgery on Pubmed Platform with Text Mining Techniques

Keywords:
Text Mining,
Text Classification,
Text Analysis,
Natural Language Processing,
Digital Data

Abstract: In this age, which is described as the digital age, both the importance and the number of digital data with the services offered by communication technologies are increasing rapidly. Text Mining (TM) methods are one of the most common methods used to extract meaningful words from complex texts and to access information. TM studies are widely used in the field of medicine. The aim of this study is to have an idea about the trend of studies in the field of surgery using TM methods from the articles published on the Pubmed platform, which is an English-language medical database. At the same time, it is to obtain keywords by using TM on the abstracts of studies in this field and to visually display the frequency values of these words. Five commonly used keywords were produced from the text by first using text pre-processing and then text analysis methods on the text file that constitutes the data set of the study. The frequency information of the generated keywords was successfully presented using graphics and word cloud.

Sorumlu Yazar/Corresponding Author: shrk19@gmail.com

1. GİRİŞ

Dijital çağ olarak nitelendirilen bu çağda içinde yaşadığımız toplumda, iletişim teknolojilerinin sunduğu hizmetlerle birlikte dijital verinin önemi de sayısı da gün geçtikte hızla artmaktadır. Durum böyle olunca ilk bakışta önemsiz ve anlamsız görülen bu dijital veri yığınları günümüz teknolojisi ve yapay zekâ (Artificial intelligence=AI) sayesinde daha anlamlı hale getirilebilmektedir. Gücünü veriden alan AI sistemleri elbette sadece sayısal ve düzenli veri ile uğraşmazlar, aynı zamanda bu teknoloji sayesinde karmaşık yapıda ve düzensiz biçimde bir araya gelmiş veriler çeşitli bilim dalları ve teknikler sayesinde daha anlamlı hale getirilebilmektedir. Bu bilim dallarından birisi de AI'nın bir kolu olan Doğal Dil İşleme (Natural Language Processing=NLP)'dir. NLP AI bir koludur ve bilgisayarların, insan dilini kavramasını, üretmesini ve idare etmesini sağlar [1].

Doğal dil belgelerini analiz etmede kullanılan Metin Madenciliği (MM) verinin anlamlı hale getirilmesinde son derece önemli bir tekniktir [2]. NLP belgelerinin anlamlı hale getirilmesi ve analiz edilmesi elbette kolay bir iş değildir. Dolayısıyla bu verilerin sınıflandırılarak değerli hale getirilmesi de epey önem kazanmıştır [3].

Tıpta üretilen dijital verilerin miktarı da bundan payını alarak önemli ölçüde bir artış göstermektedir. Dolayısıyla tıbbın farklı alanlarında da MM'nin uygulanabilirliği görülmüştür [4].

Tıpta Metin madenciliği adına yapılan çalışmalar genel olarak incelendiğinde; bir şifa ve tedavi sistemi olarak geleneksel Çin Tıbbına dair temel bilgi kaynakları MM yöntemleri ile ele alınarak gömülü verilerin açığa çıkarılması bu alanda önemli kazanımlar ortaya koymuştur. Elde edilen veriler ile hastalık ve insan yaşam sistemi hakkındaki anlayışa bütünsel bir bakış açısı sağlanmıştır [5].

Bugüne kadar veri ayıklamanın etkili bir yolu olmadığı için değerlendirilemeyen hasta tıbbi kayıtlarının MM yöntemiyle analiz edilmesine dair öneriler ve bakış açıları sunulmuştur [6].

Tarihi tıbbi belgelerdeki semantik bilgilerin sağlam tespiti için kullanılan pipeline'lar iki büyük ölçekli belge arşivine uygulanarak halka açık semantik yönelimli bir arama sistemi geliştirilmeye çalışılmıştır [7].

Yüksek karmaşıklığa sahip bir Üniversite hastanesinin veri tabanından çıkarılan ameliyatların metinsel açıklamalarını ve ameliyat sonrası hasta kayıtlarını kullanarak Cerrahi Alan Enfeksiyonlarını (CAE) tahmin etmek ve tespit etmek için MM ve makine öğrenimi yöntemlerinin kullanılabileceği önerilmiştir [8].

MM yöntemi kullanılarak, yapılandırılmamış cerrahi metin verilerini önceden işlemek için otomatikleştirilmiş entegre bir yöntem geliştirilmiştir [9].

Birden çok kaynaktan alınan klinik ve idari verileri kullanan ve cerrahi alan enfeksiyonlarının (CAE) sürveyansı için bir model geliştirilmiştir [10].

Kaya ve Gülbandılar (2022) otel yorumlarından oluşturdukları veri tabanına MM yöntemlerinden Gizli Dirichlet Ayrımı (GDA), ilişkisel konu modeli (İKM) ve yapısal konu modeli (YKM) yöntemleri kullanarak verilerden "konu tespitini" hedeflemişlerdir. Çalışmalarından YKM yönteminde daha başarılı sonuçlar elde edildiği gösterilmiştir (Tutarlılık değeri 0.509) [11].

Cengiz (2020), yapmış olduğu çalışmada hastanelerden alınan hastalık-belirti ilişkilerine dair veriler kullanılarak insanların hangi hastalığa sahip olduğuna dair en iyi sonucu bulmak için MM yöntemi kullanımı amaçlanmıştır. Kullanılan veri seti büyüdükçe alınan sonuçların daha yüksek doğruluk oranına sahip olduğunu gözlemlemiştir. Belirtilere göre hastalıklar hakkında gösterilen kısa ve öğretici bilgilerin insanların daha bilinçli hareket etmesini sağladığı ortaya koymuştur. Bu durum zamandan ve iş yükünden kazanç sağladığını ifade etmiştir [12].

Aalami (2021), yapmış olduğu çalışmada tanısal veya tarama için yapılan üst gastrointestinal sistem endoskopi raporlarını metin madenciliği yöntemi ile analiz ederek sunan yerli ve özgün bir yazılımın geliştirilmesini hedeflemiştir. Yapılan örneklemeye göre polip rapor edilmiş ve edilmemiş 148 endoskopi raporu çalışmaya dahil edilmiştir. Gaussian Naive Bayes, Multi Nomial Naive Bayes ve Logistic Regression, metotlarını kullanarak farklı sonuçlar elde edilmiştir. Logistic Regression'i metodu kullanarak elde edilen sonuçların doğruluk oranlarının %30'dan %66'ya arttığını belirlemiştir [13].

Beşkirli ve ark. (2021) sosyal medya ortamlarından biri olan Twitter metinleri üzerinde son zamanlarda aşı (vaccine) ile ilgili verilerin sayılarında önemli bir artış olduğunu tespit etmişlerdir. Çalışmalarında sosyal medya ortamındaki verileri kullanarak duygu analizleri yapmıştırlar. Bu veriler covid19 aşısı 3. faz denemeleri esnasındaki elde edilen Twitter verileri ile covid19 aşısının 3. faz sonrası seri üretim duyurusu yapıldıktan sonra elde edilen Twitter verileri analiz edilmiştir. Elde edilen sonuçlara göre aşılar hakkında veri sayısında artış ve aşı hakkındaki görüşlerin de olumlu yönde bir birikim oluşturulduğunu ortaya konulmuştur [14].

Qorib ve arkadaşları (2023) Covid 19 aşısın uygulamasına taraftar olan ve olmayan kişilerin sosyal medya platformu olan Twitter üzerinden mesajlarını analiz etmeyi hedeflemişlerdir. Bu çalışmada COVID-19 aşısı kararsızlığını analiz etmek için 3 duygu analizi hesaplama yöntemini (Azure Machine Learning, VADER ve TextBlob) kullanılmıştır. Üç vektörleştirme yönteminin (Doc2Vec, CountVectorizer ve TF-IDF) farklı kombinasyonunu kullanarak beş farklı öğrenme algoritmasını (Random Forest, Logistics Regresyon, Karar Ağacı, LinearSVC ve Naïve Bayes) tercih

etmişlerdir. Çalışmalarında Covid 19 aşısına karşı tereddütlerin zamanla azaldığını tespit etmişlerdir. Ayrıca TextBlob, TF-IDF ve LinearSVC model kombinasyonunun en iyi performansı gösterdiğini belirtmişlerdir [15].

Gowda ve arkadaşları (2022) yaptıkları çalışmada Girişimsel radyoloji (IR) araştırmalarını konu alan makalelerin çalışma türünü ve makalelerin radyoloji dışı klinisyenler tarafından farkındalığına dayalı olarak zaman içinde değerlendirmeleri belirlemeyi hedeflemişlerdir. Çalışma verilerini 1991 ve 2020 yılları arasındaki Pubmed yayınlanan radyoloji verilerini kullanmışlardır. Verilerin değerlendirmesinde klasik istatistiksel yöntemler kullanışlardır. Yıllara göre makale çalışmalarının dağılımlarını ortaya koymuşlardır [16].

Yapılan literatür taramasında, cerrahi alanda yazılmış makale özetlerine MM yöntemiyle odaklanan hiçbir çalışmaya rastlanılmamıştır. Bu çalışmanın amacı, bir tıp veri tabanı olan Pubmed platformu üzerinde bulunan ve cerrahi alan ile ilgili yayınlanmış makaleler hakkında MM yöntemi kullanılarak fikir sahibi olmaktır. Aynı zamanda bu alanda yapılmış çalışmaların özetleri üzerinde MM yöntemleri kullanılarak anahtar kelimeler elde etmek ve bu kelimelerin frekans değerlerini görsel olarak sergilemek de hedeflenirmiştir.

2. MATERYAL VE METOT

Bu çalışmada MM modellerinin geliştirilmesinde kullanılacak olan araçlar şu şekildedir;

- Python
- Jupyter Notebook

Python, metin analiz, ön ilşleme ve grafik işlemlerinde kullanılacak olan programlama dilidir.
Jupyter Notebook, elde ettiğimiz veri seti üzerinde metinleri analizi yapabileceğimiz ve python kodlarını çalıştıracağımız geliştirme ortamıdır.

Bu projede kullanılan kütüphaneler şu şekildedir:

- NLTK (Doğal dil işleme paketi)
- Matplotlib (Veri görselleştirme için)
- Seaborn (yüksek seviye görüntü arayüzü için)
- NumPy(matematiksel işlevler için)
- WordCloud (Kelime bulutu oluşturmak için)
- Re kütüphanesi (metni düzenlemek ya da metinden alt parçaları elde etmek için)

2.2. Veri Seti

Bu çalışmada İngilizce dilindeki makalelerden oluşan bir tıp veri tabanı olan Pubmed'in elektronik veri tabanına 05.12.2022 ve 06.12.2022 tarihlerinde arasında erişilerek 2021 ve 2022 yılları arasında dünya çapında cerrahi alan ile ilgili yayınlanmış makaleler taranmıştır. Taranan bu makaleler arasından 200 adet tam erişime açık makalenin özet bölümleri alınarak tek bir *"txt"* dosyası haline dönüştürülerek veri tabanı oluşturulmuştur.

2.3. Metin Ön İşleme

Bu bölümde ilk olarak düzenli ifadeler üzerinde işlem yapmamızı sağlayan "re" modülü içeri aktarılmıştır. Ardından doğal dil işlemede kullandığımız nltk kütüphanesinin omw-1.4 sürümü içeri aktarılmıştır.

Daha sonra 200 adet tam erişime açık makalenin özetlerinden oluşan ve "articles" isimli txt dosyamızı file_analyze isimli bir değişkene aktarılmıştır. Ayrıca sonuçta listelenecek olan anahtar kelimelerin gösterileceği grafikteki stün sayısını max_col komutu ile 5 adet olarak ayarlanmıştır. Metni kelimelere ayırmak için Tokenization işlemi gerçekleştirilmiştir.

Kullanıcıya sonucu daha hızlı gösterebilmek amacıyla yok sayılan "stop words" kelimeleri için nltk.corpus.stopwords.words('english') komutu kullanılmıştır. Ayrıca elde ettiğimiz sonuçlara göre benzer kökteki kelimeleri de elemek adına stpwrd.extend(new_stopwords) komutu kullanılarak bu kelimelerin belirlenip bu listeye sonradan eklenebilmesini sağlanmıştır.

Yine benzer yöntemle çeşitli imla işaretleri boşluklar ve boş paragrafları elemek için "sings" ve "signs_used_space" isimli iki ayrı liste tanımlanmıştır.

Lemmatization işlemi ile elde edilen kelimeler köklerinden ayrıştırılmıştır.

2.4. Metin Analiz

Metindeki her kelimenin sıklığını depolamak için kullanılacak olan keyword (anahtar kelime) adı verilen boş bir sözlük oluşturulmuştur.

Metindeki büyük harfler küçük harflere dönüştürülmüştür.Metindeki kelimeleri köklendirmek için kullanılan (yani, onları temel biçimlerine indirgeyen) PorterStemmer sınıfının bir örneği oluşturulmuştur. PorterStemmer örneğini kullanarak her kelimeyi köklendirilmiş daha sonra, her kelime için anahtar kelime sözlüğündeki karşılık gelen giriş artırılmış veya kelime daha önce görülmediyse sözlüğe yeni bir giriş eklenmiştir.

TF-IDF (Term Frequency - Inverse Document Frequency) ağırlıklandırma işlemi yapılmış, anahtar kelime sözlüğünü azalan sıklık sırasına göre sıralanmıştır.

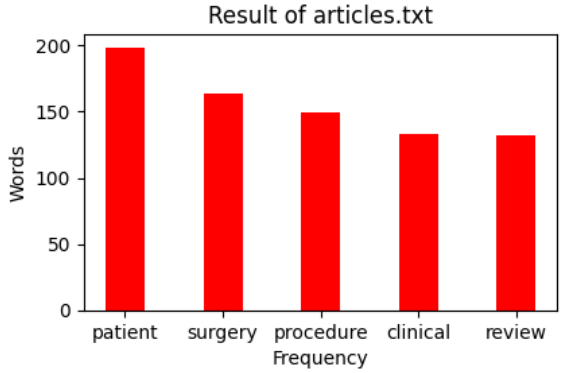
2.5. Görselleştirme ve Bilgi Edinme

Bu bölümde frekans değerleri hesaplanan kelimelerin ilk 5 tanesi "*Result of articles.txt*" başlıklı grafikte gösterilmiştir. Grafik, dosyada en sık kullanılan sözcükleri, en sık kullanılan sözcük solda ve en az kullanılan sözcük sağda olacak şekilde gösterir. Frekans

değerleri hesaplanan kelimeler bu kez "*Wordcloud*" metodu ile genişliği 3000 px ve yüksekliği 3000 olan bir kelime bulutu üzerine aktarılarak görsel hale getirilmiştir.

3. BULGULAR

Pubmed üzerinden cerrahi alanında yayınlanmış tam erişimli 200 adet makalenin özetleri alınarak tek bir txt dosyası haline getirilmiştir. Bu metin dosyası ön işlemden geçirilerek 43263 kelime yukarıda bahsedilen ön işlemlere tabi tutularak 5114 kelimeye düşürülmüştür.



Şekil1. Metinde geçen en sık 5 kelime

Metin analiz aşamasında ise Matplotlib, Seaborn ve Seaborn kütüphaneleri ile metinde sık geçen kelimeler görselleştirilmiştir. Şekil 1'de çalışma sonucunda elde edilen ve metinde en sık geçen 5 kelimenin grafiği gösterilmektedir. Buna göre; patient (hasta), surgery (ameliyat), procedure (prosedür), clinical (klinik), review (inceleme), en çok geçen kelimelerden olmuştur.

Şekil2. Frekans değerlerine göre kelime bulutu gösterimi

Şekil 2'de de yine sık geçen kelimeler sıklıklarına göre oransal olarak kelime bulutu şeklinde gösterimi verilmiştir.

4. TARTIŞMA VE SONUÇ

İçinde bulunduğumuz dijital toplumda sayısı hızla artan düzensiz verilerin işlenmesi ve aranan veriye doğru biçimde erişilmesi önemli bir araştırma konusudur. Bu doğrultuda metin madenciliği yöntemleri kullanmak bizlere oldukça fayda sağlamıştır.

Cerrahi alandaki tam erişimli 200 makalenin özetlerinin analiz edildiği bu çalışmada, literatürde yapılan çalışmalardan farklı olarak "*cerrahi*" konusuna genel olarak yaklaşılmıştır. Sonuçlara bakıldığında aynı anlama gelebilecek anahtar kelimelerin Stopword listesine eklenerek daha özel sonuçlar elde edilmiştir.

Bu çalışma yapay sinir ağları, derin öğrenme gibi tekniklerin kullanılmasıyla cerrahideki çalışmaların nereye doğru yöneldiğine dair tahmin çalışmalarına kaynak teşkil edecektir.

Etik Hususlar

Etik kurallara uyum

Bu araştırmanın planlanmasından uygulanmasına, verilerin toplanmasından verinin analizine kadar olan tüm süreçte "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir. Çalışmanın yazım sürecinde bilimsel etik ve alıntı kurallarına uyulmuş, toplanan veriler üzerinde herhangi bir tahrifat yapılmamış ve bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

Finansman
Kar amacı gütmeyen herhangi bir kuruluştan çalışma ile ilgili fon alınmamıştır.

Çıkar çatışması
Çalışma ile ilgili herhangi bir kişi veya kurumla çıkar çatışmasının bulunmadığını yazarlar olarak onaylıyoruz

KAYNAKÇA

[1] Doğal dil işleme nedir? www.ibm.com/topics/natural-language-processing (Erişim Tarihi: 16.12.2022)

[2] Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using text mining techniques for extracting information from research articles. In Intelligent natural language processing: Trends and Applications (pp. 373-397). Springer, Cham. DOI: 10.1007/978-3-319-67056-0_18

[3] Göker, H., & Tekedere, H. (2017). FATİH projesine yönelik görüşlerin metin madenciliği yöntemleri ile otomatik değerlendirilmesi. *Bilişim Teknolojileri Dergisi*, *10*(3), 291-299. DOI: 10.17671/gazibtd.331041

[4] Zhou, X., Peng, Y., & Liu, B. (2010). Text mining for traditional Chinese medical knowledge discovery: a survey. *Journal of biomedical*

informatics, *43*(4), 650-660. DOI: 10.1016/j.jbi.2010.01.002

[5] Zhou, X., Han, H., Chankai, I., Prestrud, A., & Brooks, A. (2006). Approaches to text mining for clinical medical records. In *Proceedings of the 2006 ACM symposium on Applied computing* (pp. 235-239). DOI: 10.1145/1141277.1141330

[6] Thompson, P., Batista-Navarro, R. T., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., & Ananiadou, S. (2016). Text mining the history of medicine. *PloS one*, *11*(1), e0144717. DOI: 10.1371/journal.pone.0144717

[7] da Silva, D. A., Ten Caten, C. S., Dos Santos, R. P., Fogliatto, F. S., & Hsuan, J. (2019). Predicting the occurrence of surgical site infections using text mining and machine learning. *PloS one*, *14*(12), e0226272. DOI: 10.1371/journal.pone.0226272

[8] Khaleghi, T., Murat, A., Arslanturk, S., & Davies, E. (2019). Automated surgical term clustering: A text mining approach for unstructured textual surgery descriptions. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 2107-2118. DOI: 10.1109/JBHI.2019.2956973

[9] Ciofi Degli Atti, M. L., Pecoraro, F., Piga, S., Luzi, D., & Raponi, M. (2020). Developing a surgical site infection surveillance system based on hospital unstructured clinical notes and text mining. *Surgical Infections*, *21*(8), 716-721. DOI: 10.1089/sur.2019.238

[10] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507-513. DOI: 10.1136/jamia.2009.001560

[11] Kaya, A. & Gülbandılar, E. (2022). "Konu Modelleme Yöntemlerinin Karşılaştırılması", *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, 3,(2). 46-53, DOİ:10.53608/estudambilisim.1097978

[12] Cengiz, A. (2020). "Hasta Teşhis Koyma Yardımcısı", *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, 1(2), 6-9,

[13] Aalami, N. (2021). Endoskopi Raporlarının Metin Madenciliği Algoritması Kullanılarak İncelenmesi, *Eskişehir Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği ABD*., Eskişehir.

[14] Beşkirli, A. , Gülbandılar, E. & Dağ, İ. (2021). Metin Madenciliği Yöntemleri ile Twitter Verilerinden Bilgi Keşfi. *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, 2 (1), 21-25.

[15] Miftahul Qorib, M., Oladunni, T., Denis, M., Ososanya, E. & Cota, P. (2023). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset, *Expert Systems With Applications* 212 (2023) 118715. https://doi.org/10.1016/j.eswa.2022.118715

[16] Gowda, P.C., Lobner K., Nejad N.H. & Clifford R.Weiss C.R. (2022). Bibliometric analysis of interventional radiology studies in PubMed-indexed literature from 1991 to 2020, Clinical Imaging, 85, 3-47.