Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **23** (2023) 025503 (417-426) AKU J. Sci.Eng.**23** (2023) 025503 (417-426)

DOI:10.35414/akufemubid.1206851

Araştırma Makalesi / Research Article

Veri Madenciliği Uygulamalarının Web Tabanlı Mekânsal Görsel Analitik Ortamda Sunumu: COVID-19 Aşı Tweet'leri Örneği

Burak ÇAĞLAR[1], Hüseyin Zahit SELVİ[2]

[1] *Necmettin Erbakan Üniversitesi, Fen Bilimleri Enstitüsü, Harita Mühendisliği Anabilim Dalı, Konya.*

[2] *Necmettin Erbakan Üniversitesi, Mühendislik Fakültesi, Harita Mühendisliği Bölümü, Konya.*

e-posta:bcaglar19@gmail.com, ORCID ID: http://orcid.org/0000-0002-4490-1447

hzselvi@erbakan.edu.tr, ORCID ID: http://orcid.org/0000-0001-7486-0992

Geliş Tarihi: ; Kabul Tarihi:

Anahtar kelimeler
Mekânsal görsel analitik; Veri madenciliği; Büyük veri; Etkileşimli haritalama; Sosyal medya

Öz

Mekânsal görsel analitik, mekânsal bilgilerin etkileşimli görsel ara yüzlerle ele alındığı analitik akıl yürütme bilimidir. Mekânsal görsel analitik sistemleri sayesinde, Twitter gibi sosyal medya platformlarındaki büyük veri setlerinden bir konu hakkında elde edilen veriler son kullanıcıya etkileşimli haritalama sistemleriyle sunulabilir. 11 Mart 2020'de Dünya Sağlık Örgütü'nün COVID-19 salgınını duyurmasının ardından Twitter veri trafiğinde de ciddi bir artış görülmüştür. Bu çalışmada, COVID-19 salgını döneminin önemli tartışmalarından biri olan COVID-19 aşıları hakkındaki tweet trafiğinin zamansal ve mekânsal gelişimi veri madenciliği teknikleriyle incelenmiş ve görsel analitik ortamda sunulmuştur. Bu çalışma ile twitter gibi sosyal medya platformlarının sahip olduğu büyük veri olarak kabul edilen veri setlerinin veri madenciliği yöntemleriyle analiz edilerek afet ve kriz yönetimi açısından önemli çıkarımlar yapılabileceği ortaya konmuştur.

Presentation of Data Mining Applications in Web Based Geovisual Analytical Environment: Example of COVID-19 Vaccine Tweets

Keywords
Geovisual analytic; Data mining; Big data; Interactive mapping; Social media

Abstract

Spatial visual analytics is the science of analytical reasoning in which spatial information is handled with interactive visual interfaces. Thanks to spatial visual analytics systems, data obtained from large data sets on social media platforms such as Twitter can be presented to the end user with interactive mapping systems. After the World Health Organization announced the COVID-19 outbreak on March 11, 2020, there has been a significant increase in Twitter data traffic. In this study, the temporal and spatial development of tweet traffic about COVID-19 vaccines, which is one of the important discussions of the COVID-19 epidemic period, was examined with data mining techniques and presented in a visual analytical environment. With this study, it has been revealed that important inferences can be made in terms of disaster and crisis management by analyzing the data sets, which are accepted as big data, of social media platforms such as twitter with data mining methods.

1. Giriş

Görsel analitik alanının gelişmesiyle ortaya çıkan mekânsal görsel analitik, coğrafi konum ve bu konumdaki nesneler, olaylar, olgular ve işlemleri içeren problemlerin etkileşimli görsel ara yüzlerle ele alındığı analitik akıl yürütme bilimidir (Thomas and Cook 2005, Andrienko *et al.* 2007, Andrienko *et al.* 2011). Mekânsal görsel analitik, kullanıcıların mekânsal verileri kullanarak örüntüleri tespit etmeleri ve gelecekteki sonuçları tahmin etmeleri için kartografya, hesaplamalı yöntemler, ara yüz tasarımı, biliş bilimi vb. alanların her birini etkileşimli haritalama sisteminde birleştirmektedir (Robinson 2017).

Mekânsal görsel analitik'in çeşitli alanlarda uygulamaları bulunmaktadır. MacEachren *et al.* (2011) geliştirdikleri SensePlace2 adlı uygulamada

kriz yönetimi alanında Twitter mesajlarını kullanarak coğrafi temelli durumsal farkındalık ile ilgili mekânsal görsel analitik yaklaşımı sunmuşlardır. Moore *et al.* (2013) geliştirdikleri eXplorer isimli mekânsal görsel analitik uygulamasıyla grid tabanlı toplanan WikiCrimes verileriyle analiz işlemleri gerçekleştirerek, analiz sonuçlarını kullanıcıya mekânsal olguların daha kolay anlaşıldığı nokta (pin) haritaları, yoğunluk haritaları ve koroplet haritalar ve istatistiksel bilgi grafikleri şeklinde sunmuşlardır. Luo *et al.* (2015) küresel değişim değerlendirme (global change assessment) modelinden iklim senaryolarının keşfedilmesini sağlamak için kullanıcıların su kıtlığını coğrafi farklılıklar, zamansal değişim ve gelecekteki farklı iklim politikaları senaryolarıyla karşılaştırarak keşfetmelerine olanak tanıyan web tabanlı mekânsal görsel analitik yaklaşımı sunmuşlardır. Lei *et al.* (2015) Nijer Nehri Havzasındaki su arzı, talebi ve kıtlığını çeşitli senaryolar altında keşfetmek için web tabanlı bir mekânsal görsel analitik yaklaşımı sunmuşlardır. Robinson *et al.* (2016) Suriye'deki siyasi, sosyal, ekonomik ve askeri olayların örüntülerini anlamak için mekânsal-zamansal verilerin değerlendirilmesinde STempo mekânsal görsel analitik sistem tasarımı yaklaşımını ortaya koymuşlardır.

Yakın zamanda tüm dünya'nın yaşamış olduğu COVID-19 pandemisi gibi kriz dönemlerinde; Facebook, Instagram ve Twitter gibi sosyal medya platformlarında paylaşılan haberlerin, resmi haber ajanslarından daha hızlı verilmeleri ve daha fazla içerik barındırmaları nedeniyle insanların sosyal medya platformlarında normalden daha fazla zaman geçirdikleri görülmektedir. Kullanıcı sayısı ve sosyal ifade paylaşım sınırlamalarının bulunmadığı bu tür sosyal medya platformları sürekli yapılan sosyal mesaj paylaşımlarıyla büyük verinin barındırıldığı platformlar haline dönüşmektedir. Büyük veriye dönüşen sosyal medya verileri doğru uygulama ve analiz teknikleriyle işlenerek karar vericiler için anlamlı bilgi haline dönüşmektedir (Imran *et al.* 2015, Castillo 2016, Imran *et al.* 2020). Geçmişte Pakistan Sel Felaketi (Murthy and Longwell 2012), Nepal depremi (Kumar 2020), Hindistan Sel Felaketi (Nair *et al.* 2017) vb. birçok olayda twitter verileri analiz edilerek kullanılmıştır. Eligüzel (2021), Nepal Depremiyle ilgili tweetleri kullanarak metin sınıflandırması ve konum bilgilerine yeni bir yaklaşım getirerek olası bir deprem durumunda, kapsamlı bir deprem yönetimi sağlamak için Twitter'dan faydalı bilgiler çıkarmaya odaklanmıştır. COVID-19'un küresel istatistikleri incelendiğinde, 31 Ekim 2022 tarihi itibarıyla onaylanmış COVID-19 vaka sayılarının 627 milyonu aştığı, ölüm vakalarının 6.5 milyonun üzerinde olduğu ve 12.8 milyar doz üzerinde COVID-19 aşısı yapıldığı görülmektedir (Int Kyn. 1). Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin ve Sputnik V aşılarına Dünya Sağlık Örgütü (WHO) tarafından acil kullanım onayı (EUA) verilmesiyle birlikte dünya çapında hızlı bir şekilde aşılama çalışmaları gerçekleştirilmiştir (Dayan 2021). Dünya Sağlık Örgütü'nün 11 Mart 2020'de COVID-19 salgınını ilan etmesinin ardından sosyal medya platformlarından biri olan Twitter uygulamasının kullanım trafiğinde büyük bir artış görülmüş ve İnsanların doğal yaşamlarından uzak kaldığı, sosyal medya ortamlarında vakit geçirdikleri bu dönemde pandemiye özgü birden fazla terim twitter uygulamasında trend olmuştur. Bu terimlerin en fazla kullanılanları arasında aşı ile ilgili terimler de yer almaktadır. Bu dönemdeki Twitter verileri, duygu analizi, konu modelleme, davranış analizi, veri madenciliği ve analitik görselleştirme alanlarında çalışan araştırmacılar için değerli bir kaynak haline dönüşmüştür.

Bu çalışmada, COVID-19 pandemi döneminin en büyük tartışma konularından olan COVID-19 aşıları hakkında atılan tweet'lerin hangi ülkelerin gündeminde yer aldığı ve tweet trafiğinin mekânsal – zamansal gelişimi, kümeleme analizi yöntemiyle incelenmiş ve interaktif yapılı görsel analitik ortamda sunulmuştur. Çalışmada Pfizer/Biontech, Sinopharm, Sinovac, Moderna, Oxford AstraZeneca, Covaxin, Sputnik markalarının ürettiği COVID-19 aşıları hakkında atılan tweet veri seti kullanılmıştır. İlk olarak, tweet'lerin coğrafi kodlama işlemi yapılmıştır. Bu sayede belirli bir konum bilgisi içeren tweet'lerin harita üzerinde işaretlemesi gerçekleştirilmiştir. İkinci olarak, tweet veri setinde bulunan gizli verilerin keşfi gerçekleştirilmiştir. Üçüncü aşamada, kümeleme analizi yöntemiyle birden fazla değişken kullanarak kümeleme analizi

yapılmıştır. Çalışmanın son aşamasında, COVID-19 aşısı hakkındaki tweet verilerinin mekânsal-zamansal gelişiminin gösterilmesi, kümeleme analiz sonuçlarının görsel analitik ortamda sunulması amacıyla web tabanlı interaktif bir uygulama geliştirilmiştir. Kümeleme analizi tekniğiyle, ülkelerin "covid" ve "aşı" içerikli atılan tweetler ile o ülkelerde uygulanan Covid-19 aşıları arasında mekânsal-zamansal benzerlik ve farklılıklar ortaya konulmuştur.

2. Materyal ve Metot

2.1 COVID-19 Tweet Veri Seti

Bu çalışmada, Gabriel Preda tarafından dünya genelinde Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin ve Sputnik V aşıları ile ilgili toplanarak hidratlanmış şekilde "COVID-19 All Vaccines Tweets" başlığı altında Kaggle platformunda kullanıcıların paylaşımına açılan tweet veri seti kullanılmıştır (Int Kyn. 2).

2.2 Coğrafi Kodlama

Twitter kullanıcıları, kullandıkları mobil cihazın küresel konumlandırma sistemini (GPS) kullanarak coğrafi konumlarını tweet'e eklenmesini sağlayabilmektedirler. Ancak daha önceki çalışmalar incelendiğinde;

Bennett *et al.* (2018) 2016-2017 yılları arasında Southampton şehrinde gerçekleştirdikleri çalışma kapsamında elde ettikleri 5 milyon tweet'den yaklaşık 36 bin tweet'in coğrafi konum verisine sahip olduğunu,

Burton *et al.* (2012) twitter üzerindeki sağlık iletişimi hakkında gerçekleştirdikleri çalışma kapsamında elde ettikleri 23.8 milyon tweet verisinden sadece 2.02% oranındaki yani yaklaşık 480 bin tweet'in coğrafi konuma sahip olduğunu bildirmiştir.

Birçok twitter kullanıcısı kullanıcı profillerini tanımlarken, konum bilgisi olarak şehir, eyalet, ülke isimlerini girmektedir. Ancak birçok kullanıcı da hiçbir konum ifade etmeyen rastgele ya da mizahi ifadeler girmektedir (Burton *et al.*2012).

Bu çalışmada, tweet verilerinin harita üzerinde konumlandırılması için twitter kullanıcılarının kendi konumlarını ifade ettikleri "user_location" öznitelik bilgisi kullanılmıştır. Twitter kullanıcılarının bu alana herhangi bir kısıtlama olmaksızın istediği içeriği girebildiği için öncelikli olarak yeryüzünde bir konum ifade eden içeriklerin bulunması işlemi gerçekleştirilmiştir. Konum ifade eden içeriklerin bulunması işleminde aşağıdaki hususlara dikkat edilmiştir;

- Kullanıcının tweet atarken konum bilgisini açmamış ve/veya konum kısmına herhangi bir ifade yazmamış olması (boş bırakması),
- Kullanıcının tweet atarken konum bilgisini açmış olması ya da konum bilgisini koordinat çiftleriyle (enlem/boylam) tanımlarken "°", "N, "S, "E, "W, "K, "G, "D, "B ifadelerini kullanması,
- Kullanıcının konum bilgisini kendi ana dilindeki ülke ismiyle ifade etmesi (Örneğin; الإماراتالعربيةالمتحدة),
- Kullanıcının konum bilgisini İngilizce ülke ismiyle ifade etmesi (Örneğin; United Arab Emirates),
- Kullanıcının konum bilgisini ülkelerin id/kod bilgileri ile ifade etmesi (Örneğin; AE),
- Kullanıcının konu bilgisini ülkelerin ISO3 ülke kodu bilgileri ile ifade etmesi (Örneğin; ARE).

Kullanıcı konum bilgisini yerleşim şehir ismi ile ifade etmişse bu tweetlerin bulunmasında SimpleMaps firmasının www.simplemaps.com/data/world-cities adresinde yer alan Temel düzeyde kullanıma sunulan ve öne çıkan şehirleri (büyük, başkentler vb.) içeren "csv" formatındaki veri setinden yararlanılmıştır.

2.3 Kümeleme Analizi

Kümeleme analizinin temel amacı, bir veri topluluğundaki nesneleri (birimleri) benzerliklerine göre kendi içinde olabildiğince birbirine benzeyen, kendi aralarında ise olabildiğince farklı kümelere ayırmaktır (Tatlıdil 1996, Atbaş 2008).

Nesneleri (birimleri) benzerliklerine göre kümelemede en çok kullanılan algoritmalar hiyerarşik ve hiyerarşik olmayan yöntemler olarak iki kategoriye ayrılmaktadır (Blashfield and Aldenferder 1978). Her iki yöntem de yer alan algoritmalardaki en önemli ölçüt, kümeleme analizi sonucunda oluşan kümeler arasındaki farkın ve

kümeler içi benzerliklerin maksimum düzeyde sağlanmasıdır.

Hiyerarşik kümeleme yöntemleri içerisinde en iyi kümeleme sonucunu veren yöntemler biri olan Ward tekniği, birimlerin kümelenmesinde varyansı minimuma indiren ve en uygun küme sayısını tahmin eden yöntemdir (Hands and Everit 1987). Ward yönteminde Hata Kareler Toplamı (ESS) formülünden yararlanılmaktadır.

$$ESS = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \qquad (1)$$

Burada x_i *i*'inci gözlemin skoru, *n* veri sayısıdır (Aldenderfer and Blashfield 1984).

Hiyerarşik kümeleme yöntemlerinden biri olan K-ortalamalar (K-Means) algoritması J.B. MacQueen tarafından geliştirilmiştir (MacQueen 1967). Merkez nokta kümeyi temsil etmektedir. Her veri sadece bir kümenin elemanı olması sebebiyle keskin bir kümeleme algoritmasıdır. Küme sayısının az olduğu büyük veri setlerinde çok hızlı çalışmaktadır (Han and Kamber 2001). Algoritmanın genel mantığı, *n* adet veriden oluşan veri setini, küme içi benzerliğin maksimum ve kümeler arası benzerliğinin minimum olduğu *k* adet kümeye bölümlemektir (Silahtaroğlu 2013).

Kümeleme Analizi işlemlerinde, IBM şirketinin geliştirmiş olduğu IBM SPSS (Statistical Package for the Social Sciences) Statistics yazılımı ve/veya Konstanz Üniversitesi'nde geliştirilen, açık kaynak ve çapraz platform veri analizi, raporlama, entegrasyon platform olan KNIME (Konstanz Information Miner) yazılımı kullanılmıştır.

3. Uygulama

COVID-19 pandemi döneminin en büyük tartışma konularından olan COVID-19 aşıları hakkında atılan tweet'lerin hangi ülkelerin gündeminde yer aldığı ve tweet trafiğinin mekânsal – zamansal gelişimi, kümeleme analizi yöntemiyle incelenmiş ve interaktif yapılı görsel analitik ortamda sunulması hedeflenmiştir.

Çalışma'nın ilk aşamasında, Gabriel Preda tarafından dünya genelinde Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin ve Sputnik V aşıları ile ilgili toplanarak hidratlanmış şekilde "COVID-19 All Vaccines Tweets" başlığı altında Kaggle platformunda kullanıcıların paylaşımına açılan tweet veri setinin düzenlenmesi gerçekleştirilmiştir. "csv" formatındaki "COVID-19 AllVaccinesTweets" veri seti öncelikle veri hazırlama sürecine sokularak "dbase (dbf)" formatında dönüştürülmüştür. Dönüşüm işlemi sonucunda; "tweet_id", "user_name", "user_location", "user_description", "user_created", "user_followers", "user_friends", "user_favourites", "user_verified", "date", "text", "hashtags", "source", "retweets", "favorites" ve "is_retweet" olmak üzere 16 adet öznitelik alanından ve 218412 adet tweet'ten oluşan tweet veri seti elde edilmiştir.

İkinci aşamada, COVID-19 tweet verilerinin harita ekranı üzerinde coğrafi işaretlenmesi gerçekleştirilmiştir. COVID-19 tweet veri setindeki "user_location" özniteliğinde yer alan koordinat çiftleri, anadildeki ülke isimleri, İngilizce ülke isimleri, ülke kodları ve iso3 ülke kodları sorgulamalarıyla eşleşen tweetler ve "Null" değere sahip olmayan diğer tweetler Arcgis Geocoding servisi kullanılarak harita üzerinde konumlandırma işlemi gerçekleştirilmiştir. Veri setindeki 218412 adet tweet verisinden; 142733 adet tweet harita üzerinde bir konumla eşleşmiştir. Eşleşme sağlamayan 75679 adet tweet verisinden 64348 adet tweet verisinin "user_location" özniteliği "Null" değere sahiptir. Geriye kalan 11331 adet verinin konum ifade etmeyen kelime, cümle ya da mizahi ifadeler (Örneğin; of Hope", "all around the world", "always somewhere...", "Anywhere But Here", "Around the World" vb.) içerdiği tespit edilmiştir.

Üçüncü aşamada, COVID-19 tweet veri seti içerisinde yer alan gizli örüntülerin keşfi gerçekleştirilmiştir. Örneğin; COVID-19 Tweet veri seti incelendiğinde, toplamda 218.412adet tweet olduğunu, bu tweetlerden 142.733adetinin Geocoding servisi kullanılarak coğrafi işaretleme yapıldığı bilinmektedir. Ancak bu tweetlerin hangi ülkeden atıldığı ve ülkelerde COVID19 hakkında tweet atan kullanıcı sayısının da belirlenmesi

gerekmektedir. Çalışma kapsamında aşağıdaki sorulara cevap bulmak amacıyla veri setinde gerçekleştirilen sql sorgulamaları ve mekânsal analiz teknikleriyle yeni öznitelik alanları oluşturulmuştur.

1.Tweetler hangi ülkelerden atılmıştır?

2.Bu tweetler kaç kullanıcı tarafından atılmıştır?

3.Ülkelerin tweet atan kullanıcı sayıları nelerdir?

4.Pfizer/Biontech aşısı hakkındaki tweetler hangileridir?

5.Sinopharm aşısı hakkındaki tweetler hangileridir?

6.Sinovac aşısı hakkındaki tweetler hangileridir?

7.Moderna aşısı hakkındaki tweetler hangileridir?

8.Oxford AstraZeneca aşısı hakkındaki tweetler hangileridir?

9.Covid kelimesi hakkındaki tweetler hangileridir?

10.Aşı kelimesi hakkındaki tweetler hangileridir?

Dördüncü aşamada, ülkelerin "Tweet Sayısı", "Covid Kelimesi İçeren Tweet Sayısı" ve "Vaccine (Aşı) Kelimesi İçeren Tweet Sayısı" değişkenlerine göre kümeleme analizi gerçekleştirilmiştir. Kümeleme analizinde kullanılacak veriler ilk olarak veri hazırlama sürecinden geçirilmiştir. Veri hazırlama sürecinde öncelikle veriler arasında birim farklılıkları olup olmadığı, eksik değer olup olmadığı kontrol edilmiştir. Kümeleme analizinde kullanılacak veriler aynı birimde olduğundan verilerin birimsel standardize işlemine gereksinim görülmemiştir.

Kümeleme analizi algoritmalarının verilerin birbirleriyle ilişkili olup olmadığına bakmaksızın analiz sonuçları vermesi nedeniyle kümeleme analiz işleminde önceden veriler arasındaki korelasyon testinin yapılması gerekmektedir. Birbirleriyle hiçbir ilişkisi olmayan verilerin kümelenmesi ile istenmeyen sonuçlar ortaya çıkabilmektedir (Selvi ve Çağlar 2017). Bu bakımdan veriler arasındaki korelasyonun tespiti için Pearson korelasyon testi aşağıdaki eşitliğe göre hesaplanmıştır.

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \quad (2)$$

Eşitlikte *r* iki değişken arasındaki Pearson korelasyon katsayısını, *X* ve *Y* değişkenleri, *n* veri sayısını ifade etmektedir.

Kümeleme çalışmalarında kullanılacak veriler arasındaki korelasyon katsayısının 0.80 olması gerektiği (Romesburg 1984) görüşü dikkate alındığında bu çalışmada kullanılan "Tweet Sayısı", "Covid Kelimesi İçeren Tweet Sayısı" ve "Vaccine (Aşı) Kelimesi İçeren Tweet Sayısı" değişkenlerinin birbirleriyle ilişkili olduğu gözlemlenmiştir (Çizelge 1).

Çizelge 1. Değişkenler Arasındaki Pearson Korelasyon Katsayıları

		Tweet Sayısı	Covid	Vaccine (Aşı)
Tweet Sayısı	**Pearson Korelasyon**	1	.998	.992
Covid	**Pearson Korelasyon**	.998	1	.997
Vaccine	**Pearson Korelasyon**	.992	.997	1

Veri setinin kümeleme analizi işleminde iki farklı kümeleme analizi algoritması kullanılmıştır. İlk yöntemde, her gözlemin ayrı bir küme olarak düşünüldüğü ve benzerlik katsayılarına göre birbirine en çok benzeyen kümelerin birleştirildiği yöntem olan Birleştirici Hiyerarşik Kümeleme (AGNES) algoritması Ward Tekniği kullanılarak uygulanmıştır. İkinci yöntemde ise, *n* adet veri nesnesinden oluşan bir veri kümesini, kullanıcı tarafından verilen k adet kümeye bölerek, oluşan farklı kümeler arasındaki benzerliğin minimum ve kümeler içi benzerliğin maksimum düzeyde elde edilmesini amaçlayan K-Means (K-Ortalamalar) algoritması uygulanmıştır.

Hiyerarşik kümeleme algoritması ile yapılan kümeleme işleminde dendrogram üzerinden yapılan incelemede en ideal küme sayısının 3 olduğu, K-Means algoritması ile gerçekleştirilen kümeleme analizi işleminde ise Silhoutte katsayısı testi uygulanarak en kaliteli kümeleme için k=3 olması gerektiği tespit edilmiştir. Ward yöntemiyle gerçekleştirilen hiyerarşik kümeleme işlemiyle oluşan kümelere ait bilgiler Çizelge 2'de ve K-Means algoritması ile gerçekleştirilen bölümlemeli kümeleme işlemiyle oluşan kümelere ait bilgiler Çizelge 3'te verilmiştir.

Çizelge 2. Hiyerarşik (Ward) Yöntemle Kümeleme Sonuçları

Hiyerarşik Kümeleme (Ward Tekniği)	Küme 1 (Üye:1)	Küme 2 (Üye:1)	Küme 3 (Üye:205)

Tweet Sayısı	56080	30912	9543 – 1
"Covid" kelimesi içeren Tweet Sayısı	15443	9563	2911 – 0
"Vaccine" kelimesi içeren Tweet Sayısı	22913	16383	4358 – 0

Çizelge 3.K-Means Yöntemiyle Kümeleme Sonuçları

K-Means Kümeleme	Küme 1 (Üye:2)	Küme 2 (Üye:5)	Küme 3 (Üye:200)
Tweet Sayısı	56080 – 30922	9543 - 2668	2333 – 1
"Covid" kelimesi içeren Tweet Sayısı	15443 – 9563	2911 - 706	585 – 0
"Vaccine" kelimesi içeren Tweet Sayısı	22913 – 16383	4358 - 1553	1327 – 0

Hiyerarşik kümeleme işlemi sonucunda; Hindistan 1. kümeyi, Amerika Birleşik Devletleri 2. Kümeyi ve diğer ülkeler ise 3. kümeyi oluşturmuştur. Hiyerarşik kümeleme yönteminde en yüksek tweet sayısına sahip Hindistan ve Amerika Birleşik Devletleri ayrı ayrı birer küme oluştururken, diğer ülkelerin hepsi tek bir küme de yer almıştır.

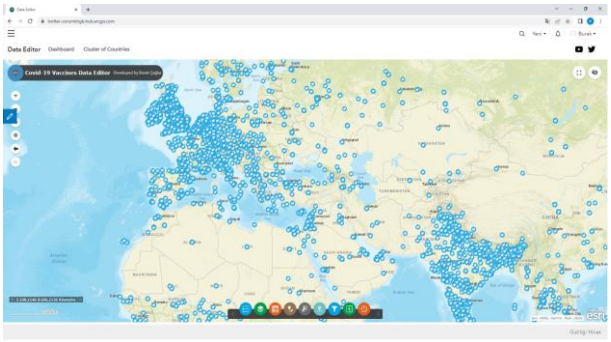
K-Means algoritması ile gerçekleştirilen kümeleme işlemi sonucunda; Hindistan ve Amerika Birleşik Devletleri 1. kümeyi, Kanada – Birleşik Krallık – Filipinler – Pakistan – Çin ülkeleri 2. kümeyi, diğer ülkeler ise 3. Kümeyi oluşturmuştur.

Çalışmanın son aşamasında, elde edilen kümeleme analiz sonuçlarının tematik haritalar şeklinde görselleştirilmesi ve analiz sonuçlarının web tabanlı interaktif uygulama üzerinde sunumu gerçekleştirilmiştir. Çalışma dünya genelindeki ülkeleri kapsadığından coğrafi işaretlemede, mekânsal analiz işlemlerinde ve tematik harita görselleştirmelerinde 1:1 milyon ölçekli Ülke İdari Sınır verileri (Int Kyn. 3) Web Mercator projeksiyonunda kullanılmıştır. Tematik harita renk seçimlerinde kartografik tasarım içim renk seçimi tavsiyesi veren çevrimiçi Color Brewer yazılımı kullanılmıştır (Int Kyn. 4).

ArcGIS Pro 2.9 yazılımı kullanılarak hazırlanan tematik haritalar ve öznitelik tabloları, Web üzerinde yayınlanmak amacıyla ArcGIS Online sunucuları üzerine Veri Düzenleme, Haritalama ve Sorgu özellikleri ile servis edilmiştir. ArcGIS Online sunucuları üzerine servis edilen Web Haritaları; ArcGIS Web App Builder, ArcGIS Dashboard, ArcGISHub yazılımları kullanılarak interaktif web uygulaması geliştirilmiştir. Bu uygulama üzerinden coğrafi veriye ve öznitelik bilgilerine erişim sağlanabilmektedir. Tasarımcı tarafından oluşturulan yapılara göre sorgulama ve analizler gerçekleştirilmekte, veri tabanında yer alan bilgilere çeşitli grafik araçlarla ve haritalar üzerindeki açılır pencereler ile interaktif şekilde erişim sağlanabilmektedir.

Geliştirilen web uygulamasına https://twitter-corumkhgb.hub.arcgis.com/ internet adresinden üzerinden kullanıcı bilgileri ile erişim sağlanmaktadır. Uygulama "Data Editor", "Dashboard" ve "Cluster of Countries" şeklinde isimlendirilen üç sekmeden oluşmaktadır.

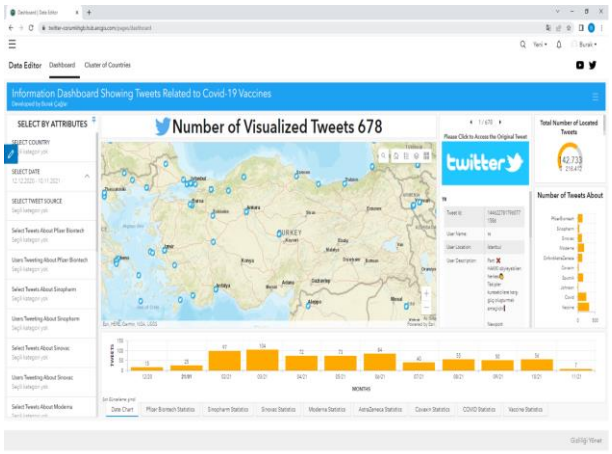
"Data Editor" olarak adlandırılan ilk sekmede, COVID-19 veri setindeki verilerin harita üzerindeki dağılımları twitter sembolüyle görülmektedir (Şekil 1). Bu sayfa üzerinde verilerin öznitelik bilgileri görüntülenmekte ve düzenlenebilmektedir. Veriler üzerinde zamansal filtreleme işlemi gerçekleştirilebilmekte, verilerin mekânsal-zamansal gelişimi izlenebilmekte ve farklı katmanlar arasından swipe (sıyırma) aracıyla karşılaştırma işlemleri gerçekleştirilmektedir.



Şekil 1. Web Uygulaması "Data Editor" Sekmesi

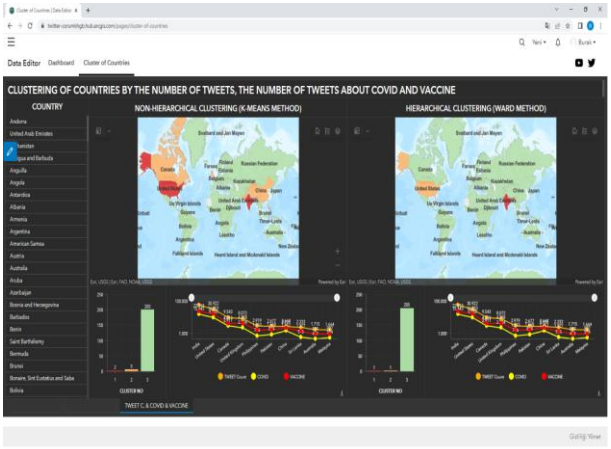
"Dashboard" olarak adlandırılan ikinci sekme, COVID-19 veri setindeki verilerden mekânsal analiz işlemleri sonucunda elde edilen bilgilerin gösterildiği paneldir (Şekil 2). Burada, ilk sekmede olduğu gibi verilerin harita üzerinde dağılımları görülmekte, ayrıca tweetlerin mekânsal – zamansal

dağılımları sütun grafik üzerinden aylık bazlı olarak takip edilebilmektedir. Harita penceresinin üzerindeki pencerede o an ekran yayılımın da olan tweet sayısı görülmekte, harita ekranının sağındaki pencerede yayılımda olan veya seçimi yapılan tweet'e ilişkin öznitelik bilgileri görüntülenmekte ve Twitter logosuna tıklandığında ise o tweet'e ilişkin orijinal twitter sayfasına erişim sağlanmaktadır. Harita penceresinin solundaki pencerede ön tanımı yapılmış çeşitli öznitelik sorgulamaları yapılabilmekte ve sorgu sonuçları harita üzerinde görüntülenebilmektedir. Zamansal değişimi gösteren sütun grafiğin altındaki sekmelerden Pfizer/Biontech, Sinopharm, Sinovac, Moderna, AstraZeneca, Covaxin aşıları ile Covid ve Aşı kelimelerini içeren tweetlere ilişkin ülkesel bazlı istatistikî bilgilere erişilmektedir.



Şekil 2. Web Uygulaması "Dashboard" Sekmesi

"Cluster of Countries" olarak adlandırılan son sekmede, ülkelerin "Tweet Sayısı", "Covid Kelimesi İçeren Tweet Sayısı" ve "Vaccine (Aşı) Kelimesi İçeren Tweet Sayısı" değişkenlerine göre hiyerarşik kümeleme ve K-Means kümeleme algoritmaları ile gerçekleştirilen kümeleme analizi sonuçlarına dayalı tematik görselleştirmeler yan yana gösterilmektedir (Şekil 3). Harita penceresinin altında yer alan sütun grafikte küme bilgileri, sütun grafiğin yanında yer alan çizgi grafikte ise harita ekranında yer alan ülkelerin kümeleme işleminde kullanılan öznitelik değişkenlerine ait değerler görülmektedir. Harita penceresinin sol kısmında ise ülkeler listesi yer almaktadır. Bu pencerelerde birbirleriyle etkileşimli olarak çalışmaktadır.



Şekil 3. Web Uygulaması "Cluster of Countries" Sekmesi

4. Araştırma Bulguları

COVID-19 pandemi döneminin en büyük tartışma konularından olan COVID-19 aşıları hakkında atılan tweet'lerin hangi ülkelerin gündeminde yer aldığı ve tweet trafiğinin mekânsal – zamansal gelişimi, kümeleme analizi yöntemiyle incelenmiş ve sonuçlar interaktif yapılı görsel analitik ortamda sunulmuştur.

COVID-19 Tweet veri setinde bulunan toplam 218.412 adet tweet verisinden; 142.733 adet tweet harita üzerinde coğrafi olarak işaretlenmiştir. Harita üzerinden coğrafi işaretlemesi gerçekleştirilen 142.733 adet tweet dünya genelinde 207 ülkeden atılmıştır. Atılan tweet sayılarına bakıldığında; Hindistan 56.080 tweetle 1.sırada, Amerika Birleşik Devletleri 30.922 tweetle 2.sırada ve Kanada 9.543 tweetle 3.sırada yer almıştır.

Covid-19 Tweet veri seti twitter kullanıcı bazlı incelendiğinde; 218.412 adet tweet'in toplam 86.558 farklı kullanıcı tarafından atıldığı, 12.210 adet tweet ile Hindistan'dan "CoWIN Clore 18-44" isimli kullanıcının 1.sırada, 11.495 adet tweet ile konum bilgisi bulunmayan "CowinBangalore" isimli kullanıcının 2.sırada ve 5.710 adet tweet ile yine Hindistan'dan "VaxBLR" isimli kullanıcının olduğu tespit edilmiştir.

Covid-19 Tweet veri seti atılan tweet'lerin zamansal gelişimi açısından incelendiğinde; veri setinin 12 Aralık 2020 ile 10 Kasım 2021 tarihi aralığında atılan tweetleri kapsadığı, 12 Şubat 2021 tarihine kadar atılan tweet sayısının 6.986 olduğu, Mart/2021

tarihinde atılan tweet sayılarının hızlı bir yükselişe geçtiği ve Nisan/2021 ayı itibarıyla aylık ortalama 25 bin tweet'in atıldığı görülmüştür.

Covid-19 Tweet veri seti atılan tweet'lerin kaynağı (kullanılan platform açısından) incelendiğinde; 218.412 adet Tweet'in toplam 369 farklı kaynaktan atıldığı, 56.496 adet tweet ile "Twitter for Android" kaynağının 1. sırada, 54.363 adet tweet ile "Twitter Web App" kaynağının 2.sırada ve 46.614 adet tweet ile "Twitter for iPhone" kaynağının 3.sırada olduğu belirlenmiştir.

Covid-19 Tweet veri seti analiz edildiğinde; Covaxin aşısı hakkında 69.844 adet tweet (konumlu 37.700 / konumsuz 32.144), Moderna aşısı hakkında 46.007 adet tweet (konumlu 31.711 / konumsuz 14.296), Pfizer/biontech aşısı hakkında 21.484 adet tweet (konumlu 14.232 / konumsuz 7.252), Sputnik aşısı hakkında 17.703 adet tweet (konumlu 12.713 / konumsuz 4.990), Sinovac açısı hakkında 10.960adet tweet (konumlu 7.433 / konumsuz 3.527), Sinopharm aşısı hakkında 8.867 adet tweet (konumlu 6.588 / konumsuz 2.279), AstraZeneca aşısı hakkında 8.325 adet tweet (konumlu 5.712/ konumsuz 2.613) ve Johnson aşısı hakkında 2.147 adet tweet (konumlu 1.383 / konumsuz 764) tespit edilmiştir. Ayrıca "Covid" kelimesi içeren 55.229 adet tweet'in (konumlu 41.530 / konumsuz 13.699) ve "Vaccine / aşı" kelimesi içeren 96.299 adet tweet'in (konumlu 69.608 / konumsuz 26.691) olduğu tespit edilmiştir.

Veri madenciliği tekniklerinden biri olan kümeleme analizi yöntemiyle ülkelerin "Tweet Sayısı", "Covid Kelimesi İçeren Tweet Sayısı" ve "Vaccine (Aşı) Kelimesi İçeren Tweet Sayısı" değişkenleri ile gerçekleştirilen;

Hiyerarşik kümeleme işlemi sonucunda, Hindistan 1. kümeyi, Amerika Birleşik Devletleri 2. kümeyi ve diğer ülkeler (205 ülke) ise 3. kümeyi oluşturmuştur. Hiyerarşik kümeleme yönteminde en yüksek tweet sayısına sahip Hindistan ve Amerika Birleşik Devletleri ayrı ayrı birer küme oluştururken, diğer ülkelerin hepsi tek bir küme de yer almıştır.

K-Means algoritması ile gerçekleştirilen kümeleme işlemi sonucunda ise; Hindistan ve Amerika Birleşik Devletleri 1. kümeyi, Kanada – Birleşik Krallık – Filipinler – Pakistan – Çin ülkeleri 2. kümeyi, diğer ülkeler ise 3. kümeyi oluşturmuştur.

5. Sonuçlar ve Öneriler

Sosyal medya'nın kullanımı arttıkça insanlar arasındaki iletişim kurma şeklide değişmektedir. Toplumum tamamını ilgilendiren pandemi gibi acil durumlarda, kriz yöneticilerinin olaya uygun şekilde müdahale etmelerine yönelik veriler sosyal medya aracılığıyla kullanıcılardan toplanabilir. Sosyal medya sayesinde, meydana gelen bir olaya ilişkin gerçek zamanlı bildirimler analiz edilerek, yaşanılan olaya ilişkin etkilerin ilk farkındalıkları ve olaya ilişkin toplumun tepkisi belirlenebilir.

Statista firması tarafından yayınlanan "Ocak 2022 İtibarıyla Twitter Kullanıcı Sayısına Göre Önde Gelen Ülkeler" raporunda; Amerika Birleşik Devletleri 76.9 milyon kullanıcı ile 1.sırada, Japonya 58.95 milyon kullanıcı ile 2. sırada ve Hindistan'ın 23.6 milyon kullanıcı ile 3. sırada olduğu görülmektedir (Int Kyn. 5). Bu çalışma kapsamında elde edilen analiz sonuçları incelendiğinde, Hindistan 56.080 tweetle 1.sırada, Amerika Birleşik Devletleri 30.922 tweetle 2.sırada ve Kanada 9.543 tweetle 3.sırada yer aldığı görülmektedir. Twitter kullanıcı sayısı bakımından 58.95 milyon kullanıcı ile 2.sırada bulunan Japonya'dan Covid-19 aşısı hakkında toplamda 564 adet tweet atılırken, Twitter kullanıcı sayısı bakımından 7.9 milyon kullanıcı sayısı ile dünya genelinde 14. sırada yer alan Kanada'nın Covid-19 aşısı hakkında toplamda attığı 9.543 tweetle 3.sırada yer alması dikkat çeken bir sonuç olmuştur. Ülkelerin Covid-19 aşılarına verdikleri ön kullanım onayları ve gerçekleştirdikleri aşılama çalışmaları dikkate alındığında, Kanada da Pfizer/Biontech aşısının toplu aşılama çalışmalarına 14 Aralık 2020'de başlanılırken (Int Kyn. 6), Japonya'da Pfizer/Biontech aşısının toplu aşılama çalışmalarına 17 Şubat 2021'de başlanılmış (Int Kyn. 7) olması, ülkelerden atılan Covid-19 aşısı hakkındaki tweet'ler ile ülkelerin gerçekleştirmiş oldukları aşılama takvimi arasında doğrusal bir ilişki bulunduğu şeklinde yorumlanmıştır. Yine aynı şekilde Covid-19 veri setindeki tweet trafiğinde yaşanılan artışların ülkelerin aşılama takvimleriyle ilişkili olduğu

gözlemlenmiştir. Ancak atılan tweet'lerin Covid-19 aşılarını destekler nitelikte mi olduğu yoksa Covid-19 aşısı karşıtlığı hakkında mı olduğu bu çalışma kapsamında gerçekleştirilmemiştir. Bu konu, twitter mesajlarından duygu analizi tespiti kapsamında başka çalışmalar içeriğinde gerçekleştirilebilir.

Bu çalışmada kullanılan Covid-19 Tweet veri seti'nin kümeleme analizinde, K-Means algoritmasının Hiyerarşik kümeleme algoritmasına göre daha iyi bir kümeleme analizi yaptığı değerlendirilmiştir. Çok değişkenli kümeleme analizi sonucunda oluşan kümeler incelendiğinde, Hindistan ve Amerika Birleşik Devletleri ülkelerinden atılan tweetlerin yüksek düzeyde "covid" ve "Vaccine/aşı" kelimelerini içererek aynı kümede bulundukları görülmektedir. Yine Kanada, Birleşik Krallık, Filipinler, Pakistan ve Çin ülkelerinden atılan tweetlerinde aynı düzeyde "Covid" ve "Vaccine/aşı" kelimeleriyle ilgili oldukları ve ülkelerin aynı grupta oldukları gözlemlenmektedir.

Burada verilen sonuçlar yürütülmekte olan doktora tez çalışmasının ilk bulgularıdır. Bundan sonraki çalışmada, ülkelerin uygulamış oldukları aşı türleri ve dozları hakkındaki bilgiler araştırılıp uygulamaya dahil edilerek çalışma kapsamında elde edilen sonuçların karşılaştırılması ve desteklenmesi planlamaktadır.

6. Kaynaklar

Aldenderfer, M.S., R.K. Blashfıeld, 1984. Cluster analysis, Beverly hills: Sage Publications.

Andrienko, G., Andrienko, N., Jankowski, P., Keim, D., Kraak, M. J., MacEachren, A. M., and Wrobel, S.,2007. Geovisual analytics for spatial decision support: setting the research agenda. *International Journal of Geographical Information Science*, **21(8)**, 839-857.

Andrienko, G., Andrienko, N., Keim, D., MacEachren, A. and Wrobel, S., 2011. Challenging problems of geospatial visual analytics, *Journal of Visual Languages and Computing*, **22 (4)**, 251-256.

Atbaş, A.C.G., 2008. Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara.

Bennett, N.C., Millard, D.E., Martin, D., 2018. Assessing twitter geocoding resolution. In: Proceedings of the 10th ACM Conference on Web Science, 239–243.

Blashfield, R.K., Aldenferder, M.S., 1978. The literature on cluster analysis, *Multivariate Behavioral Research*, **13**, 271-295.

Burton, S.H., Tanner, K.W., Giraud-Carrier,C.G., West, J.H., Barnes, M.D., 2012. "right time, right place" health communication on twitter:value and accuracy of location information. *Journal of Medical Internet Research*, **14(6), 1-11.**

Castillo, C., 2016. Big crisis data: Social media in disasters and time-critical situations. Cambridge University Press.

Dayan, S., 2021. COVID-19 ve Aşı, *Dicle Tıp Dergisi / Dicle Medical Journal*, **48** (Özel Sayı / Special Issue): 98-113.

Eligüzel, N., 2021. Using twitter for situational awareness after an earthquake: The roles of text categorization and location information, Doktora Tezi, Gaziantep Üniversitesi Fen Bilimleri Enstitüsü, Gaziantep.

Han, J., Kamber, M., 2001. "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers Inc.

Hands, S., Everit, B., 1987. A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical cluster techniques. *Multivariate Behaviral Research*, **22**, 235-243.

Imran, M., Castillo, C., Diaz, F., Vieweg, S., 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, **47(4)**, 1–38.

Imran,M., Mitra, P., Castillo, C., 2016.Twitter as a lifeline: Human annotated twitter corpora for nlp of crisis-related messages. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA): Paris,France.

Imran, M., Ofli, F., Caragea, D., Torralba, A., 2020. Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions. *Information Processing & Management*, **57**, 1-9.

Kumar, P.,2020. Twitter, disaster and cultural heritage: A case study of the 2015 Nepal earthquake, *Journal of Contingencies and Crisis Management*, **28**, 453–465.

Lei, T., Liang, X., Mascaro, G., Luo, W., White, D., Westerhoff, P., and Maciejewski R., 2015. An Interactive Web-Based Geovisual Analytics Tool to Explore Water Scarcity in Niger River Basin, Workshop on Visualisation in Environmental Sciences (EnvirVis).

Luo, W., Chang, Z., Kong, L.L., Link, R., Hejazi, M., Clarke, L., and Maciejewski, R., 2015. Web-Based Visualization of the Global Change Assessment Model. In: Proceedings of Visualization in Environmental Sciences (EnvirVis 2015), EuroVis 2015. Cagliari, Italy: May 25-26.

MacEachren, A.M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., and Blanford, J., 2011. SensePlace2: Geotwitter Analytics Support for Situation Awareness. 2nd IEEE Conference on Visual Analytics Science and Technology 2011, VAST 2011- Providence, RI, United States, Pages 181-190.

Moore, A., De Oliviera, M., Caminha, C., Furtado, V., Basso, V. and Ayres, L., 2013. Applying Geovisual Analytics to Volunteered Crime Data, Geospatial Visualisation, Lecture Notes in Geoinformation and Cartography, Springer-Verlag Berlin Heidelberg.

Murthy, B. and Longwell S.A., 2012. Twitter And Disasters, *Information Communication&Society*, **16(6)**, 1-19.

Nair, M.R., Ramya, G.R.,Sivakumar, P.B., 2017. Usage and analysis of Twitter during 2015 Chennai flood towards, disaster management, *Procedia Computer Science*, **115** ,350–358.

Robinson, A.C., 2017. Geovisual Analytics, *The Geographic Information Science&Technology Body of Knowledge* (3rd Quarter 2017 Edition).

Robinson, A. C., Peuquet, D. J., Pezanowski, S., Hardisty, F. A., and Swedberg, B., 2016. Design and evaluation of a geovisual analytics system for uncovering patterns in spatio-temporal event data. *Cartography and Geographic Information Science*, 1-13.

Romesburg, H.C., 1984. Cluster Analysis for Researchers, Belmont, CA: Lifetime Learning Publications.

Selvi, H.Z., Çağlar, B., 2017. Çok Değişkenli Haritalama İçin Kümeleme Yöntemlerinin Kullanılması, *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi***, 6(2)**, 415-429.

Silahtaroğlu, G., 2013. Veri Madenciliği (Kavram ve Algoritmaları), Papatya Yayıncılık, İstanbul.

Tatlıdil, H., 1996. Uygulamalı Çok Değişkenli İstatistiksel Analiz, Hacettepe Taş. Yayınları, Ankara.

Thomas, J. and Cook, K., 2005. Illuminating the Path: Research and Development Agenda for Visual Analytics, IEEE Press, 194 p.

İnternet kaynakları

1- https://covid19.who.int/ (31.10.2022)

2- https://www.kaggle.com/datasets/gpreda/all-covid19-vaccines-tweets (06.10.2021)

3- https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/countries (01.03.2021)

4- https://colorbrewer2.org/ (20.03.2021)

5- https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (31.10.2022)

6- https://en.wikipedia.org/wiki/COVID-19_vaccination_in_Canada (31.10.2022)

7- https://en.wikipedia.org/wiki/COVID-19_vaccination_in_Japan (31.10.2022)