International Journal of Social Inquiry *15*(1), 2022, pp. 55–68
journal homepage: https://dergipark.org.tr/en/pub/ijsi

International Journal of Social Inquiry
International Peer-Reviewed Open Access Electronic Journal
E-ISSN: 1307-9999
Volume 15 Issue 1 June 2022

RESEARCH ARTICLE / Araştırma Makalesi
https://doi.org/10.37093/ijsi.928685

Sosyal Medyada Duygu Analizi: COVID-19 Sürecinde 5G Algısı

Elçin Timur Çakmak[*] Ayşe Oğuzlar[**]

Öz

Bu çalışmada toplum için fırsatlar yaratacak yeni yetenekler getirmesi beklenen beşinci nesil hücresel ağlar (5G) ile COVID-19 aşısının dünya genelinde insanlar üzerinde oluşturduğu algının Duygu Analizi yöntemi ile ölçülmesi hedeflenmektedir. Bu amaçla, yaygın olarak kullanılan bir sosyal medya aracı olan Twitter'dan Ekim – Aralık 2020 tarihleri arasında 25642 adet tweet çekilmiş ve Python yazılımı aracılığı ile hesaplamalar yapılmıştır. Buna göre dünya genelinde Twitter üzerinden fikrini beyan eden kişilerin %36,4'ünün 5G ile COVID-19 aşısı hakkında pozitif algıya sahip olduğu görülmüştür. Tweet atan kişilerin %35,6'sının ise konuyla ilgili olarak pozitif ya da negatif görüşe sahip olmadığı ve %28'inin de negatif görüş bildirdiği sonucuna varılmıştır. Tüm tweetler için genel duygu skoru ortalaması 0,15 olarak bulunmuştur. Çalışmada ayrıca verilere makine öğrenmesi yöntemlerinden Sınıflandırma ve Regresyon Ağaçları (CART), Naïve Bayes (NB), k-En Yakın Komşuluk (KNN) ve Rastgele Orman (RF) algoritmaları uygulanmıştır. Elde edilen bulgulara göre sınıflandırmada en iyi sonuçları 0,7852 kesinlik (P) ve 0,7445 doğruluk (A) değerleri ile NB; 0,8209 duyarlılık (R) değeri ile KNN ve 0,7866 F-ölçütü (F) değeri ile RF algoritmaları vermiştir.

Anahtar Kelimeler: 5G, COVID-19 aşısı, duygu analizi, makine öğrenmesi, sınıflandırma algoritmaları

Jel Kodları: C11, C38, C45.

Cite this article: Timur Çakmak, E., & Oğuzlar, A. (2022). Sosyal medyada duygu analizi: COVID-19 sürecinde 5G algısı *International Journal of Social Inquiry 15*(1), 55–68. https://doi.org/10.37093/ijsi.928685

[*] Bursa Uludağ Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Bölümü, İstatistik ABD, Görükle Kampüsü, Bursa / Türkiye. Sorumlu Yazar. E-posta: elcintimur@gmail.com, ORCID: https://orcid.org/0000-0003-3247-6823

[**] Prof. Dr., Bursa Uludağ Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Ekonometri Bölümü, İstatistik ABD, Görükle Kampüsü, Bursa / Türkiye. E-posta: ayseog@uludag.edu.tr, ORCID: https://orcid.org/0000-0003-3228-9366

Article Information
Received 27 April 2021; Revised 30 July 2021; Accepted: 28 Feb 2022; Available online: 30 June 2022

Sentiment Analysis on Social Media: The 5G Perception During COVID-19 Pandemic

Abstract

This study is aimed at measuring the perception created by the COVID-19 vaccines on people around the world using Sentiment Analysis – taking into account the fifth generation (5G) cellular networks, which are expected to bring new capabilities that will create opportunities for society. For this purpose, 25642 tweets were taken from Twitter between October and December 2020 and analyzed using Python software. Accordingly, 36.4% of people worldwide who expressed their opinions on Twitter have a positive perception of 5G and the COVID-19 vaccine, also 35.6% of the tweeters do not have a positive or negative view (neutral) has been observed. However, it was observed that 28% of the people expressed negative opinions. The overall sentiment score is 0.15. Also, in this study, Classification and Regression Trees (CART), Naïve Bayes (NB), k-Nearest Neighbour (KNN), and Random Forest (RF) algorithms were applied. According to the findings, the best results were obtained by NB with 0,7852 precision (P) and 0,7445 accuracy (A) values, KNN with 0,8209 recall (R) value, and RF with 0,7866 F-measure (F) value.

Keywords: 5G, COVID-19 vaccine, sentiment analysis, machine learning, classification algorithms.

Jel Codes: C11, C38, C45.

1. Giriş

Son yıllarda sosyal medyanın hayatımıza girmesine paralel olarak bu mecralardan yararlanan kullanıcı sayısında gözle görülür bir artış yaşanmaktadır. Gerek bilgisayar ve tabletlerden gerekse cep telefonlarından Wi-Fi ve mobil internet bağlantıları aracılığıyla kullanıcılar istedikleri her an her türlü ortamdan sosyal medyaya erişim sağlayabilmektedir. Bu erişimler vasıtasıyla da kullanıcılar sosyal medya üzerinden görüş bildirmek amacıyla istedikleri herhangi bir konuda fikirlerini beyan edebilmektedirler. Kullanıcıların bu amaç doğrultusunda yoğun olarak kullandıkları sosyal medya ortamlarından biri Twitter'dır. Bu mecraya giriş yaparak tweet atan kullanıcıların bildirdikleri görüşlerin toplu olarak ilk bakışta ne yönlü olduğunu anlamak mümkün değildir. Bu nedenle bu metinlerde yer alan duygunun ne olduğunu anlayabilmek amacıyla Duygu Analizi tekniği kullanılmaktadır.

Bu çalışmada, 5G ağ teknolojisi ile COVID-19 aşısı arasında var olabilecek bir ilişkinin dünya genelindeki insanlar üzerinde nasıl bir algı oluşturduğu Duygu Analizi yöntemi ile araştırılmaktadır. 5G terimi, beşinci nesil mobil ağ teknolojisine karşılık olarak kullanılmaktadır. 1G, 2G, 3G ve 4G ağlarından sonra kullanılmaya başlanması planlanan yeni bir küresel kablosuz ağ standardıdır. 5G; makineler, nesneler ve cihazlar dahil olmak üzere hemen hemen herkesi ve her şeyi birbirine bağlamak için tasarlanmış yeni bir ağ türü sağlamaktadır.

5G akıllı şehirlerde, içinde yaşayan insanların hayatlarını daha verimli hale dönüştürmek için çeşitli şekillerde kullanılabilir. Öncelikle insanlar ve nesneler arasında daha fazla bağlantı kurmak, daha yüksek veri hızı elde etmek, otomotiv güvenliği sağlamak ve altyapı geliştirmek gibi alanlarda her zamankinden daha düşük gecikme süresi gibi daha büyük verimlilikler sağlayabilmektedir (Qualcomm, t.y.). Sağladığı verimliliğin yanı sıra kullandığı yüksek frekans dalgalarının önceki nesil ağ teknolojilerine kıyasla daha yüksek olmasından dolayı 5G'nin insan sağlığına negatif etkileri olabileceğinden de söz edilmektedir.

Tüm cep telefonu teknolojilerinin kullandığı elektromanyetik radyasyon, bazı kişilerin belirli kanser türlerinin gelişmesi de dahil olmak üzere artan sağlık riskleri konusunda endişelenmesine neden olmaktadır (Reality Check Team, 2019). 5G teknolojisinin yaydığı yüksek frekans dalgalarının insan sağlığına zararlı olmadığı yönünde bir görüş de mevcuttur.

Bu nedenle, COVID-19 pandemi döneminde tekrar tartışma konusu haline gelen 5G teknolojisinin insanlar üzerindeki etkisi araştırma konusu olarak ele alınmıştır.

Literatür incelendiğinde; son dönemlerde Twitter üzerinden yapılan güncel Duygu Analizi ve Makine Öğrenmesi çalışmaları dikkat çekmektedir. Rajput vd. (2020), COVID-19 pandemisine ilişkin Twitter'dan çektikleri veriler üzerinden unigram, bigram and trigram frekanslarını hesaplayarak modelleme yapmışlardır. Adwan vd. (2020), Twitter verileri üzerinden Duygu Analizi yapmak için kullanılan algoritmalar ve yaklaşımlar hakkında bir literatür çalışması yapmışlardır. Mehta ve Pandya (2020), farklı makine öğrenmesi ve sözlük tabanlı araştırma yöntemleri kullanarak bunun neden olduğu etkileri incelemişlerdir. Kumar vd. (2020), Malayalam dilinin kapsamını görebilmek amacıyla Twitter verilerinden yola çıkarak Duygu Analizi çalışması yapmış, Maksimum Entropi ve Naïve Bayes algoritmalarını kullanarak da makine öğrenmesi çalışması yapmışlardır. Anvar Shathik ve Krishna Prasad (2020), Duygu Analizi ve Makine Öğrenimi teknikleri için yaygın araştırma tekniklerini ve uygulamalarını analitik olarak sınıflandırmış ve analiz etmişlerdir. Poria vd. (2020) yayınladıkları makalede, gerçek duygu anlayışına ulaşmak için bu alanın eksikliklerine ve yeterince araştırılmamış temel yönlerine işaret ederek bu algıyı tartıştıklarını ve önemli sıçramaları analiz ettiklerini belirtmişlerdir. Ayrıca, bu alan için gözden kaçan ve cevaplanmayan birçok soruyu kapsayan olası bir rota çizmeye çalıştıklarını ifade etmişlerdir. Piksina ve Vernholmen (2020) ise yayınladıkları çalışmada, COVID-19 salgını sırasında koronavirüs ile ilgili algının İsveç borsa getirileri üzerindeki etkisini incelemişlerdir. Bahja ve Safdar (2020) yaptıkları çalışmada, COVID-19'u 5G'ye bağlayan tweetlerin NLP tabanlı analizi sunmuşlardır. Tweetlerin analizi ve konuların belirlenmesi için duygu analizi ve çeşitli teknikler içeren Doğal Dil İşleme modelleri uygulanmıştır. Wilson ve Wiysonge (2020) koronavirüs ile ilgili yaptıkları çalışmada, sosyal medyadaki örgütlenme ile aşı güvenliği konusunda halkın şüpheleri arasında önemli bir ilişki olduğu sonucuna varmışlardır. Ayrıca, karalama kampanyaları ile azalan aşı kapsamı arasında önemli bir ilişki olduğunu da vurgulamışlardır. Bužić (2019) yaptığı çalışmada; makine öğrenimi, sözlük tabanlı yaklaşım, veri sınıflandırması ve derin öğrenme gibi mevcut araştırma yaklaşımlarına odaklanan duyarlılık analizi alanında literatüre genel bir bakış sunmuştur. Duygu analizinde nispeten yeni bir alt alan olan duygu madenciliği yaptığı çalışmada tartışma konusu olmuştur. Tian vd. (2018), duygu skorunu oluşturan iki durum olan polarite ve insensite kavramlarını incelemişlerdir. Timur Çakmak ve Oğuzlar (2020), çektikleri tweetleri çalışmaya uygun hale gelecek şekilde ön işleme aşamasından geçirerek elde edilen veriler üzerinden duygu skorlarını hesaplamış ve makine öğrenmesi aşamasında sınıflandırma tekniklerinden Naïve Bayes tekniğini kullanarak model başarım ölçütlerinden doğruluk kriterine göre kıyaslama yapmışlardır.

2. Materyal ve Yöntem

Bu çalışmada, Duygu Analizi tekniği kullanılmıştır. Duygu analizi; insanların fikirlerini, duygularını, tutumlarını konulara ve olaylara bağlı olarak inceleyen bir çalışma alanıdır (Ha vd., 2019, s. 9). Analizin temelini oluşturan duygu kavramı ise, bir duruma veya olaya ilişkin bir bakış açısı veya ona yönelik bir tutum olarak tanımlanabilir (Kim & Jeong, 2019, s. 32). Duygu, insan dilinde aktarılan önemli bir bilgi türüdür (Tian vd., 2018, s. 40). Duygular; yargılar, içgörü veya insanların görüşleri aracılığıyla çeşitli hareketlerle ifade edilebilir. Bir duygu, duruma göre kişinin bilinçli veya bilinçsiz olarak ani tepki vermesi olarak ifade edilebilir. Duygu metin biçiminde iki farklı şekilde incelenebilir. Birincisi yazar üzerinde bıraktığı etki, yani yazarın duygularını ifade etmek için seçtiği kelimeler şeklinde ifade edilebilir. İkincisi ise okur üzerinde bıraktığı etki, bu

durumda da okurun yazıyı yorumlama kabiliyeti ve anlık duygu durumuna göre okuduğunu algılama biçiminde olabilmektedir (Mehta & Pandya, 2020, s. 601).

Duygu Analizi, yapılandırılmamış içerikteki duyguyu anlamaya yarayan bir araştırma alanıdır (Poria vd., 2020, s. 1). İnsanlar geçmişte arkadaşlarının ve akrabalarının tavsiyelerine güveniyorlardı; ancak bu tavsiyeler kişilerle sınırlı olmaktaydı. Bugün ise hiç tanımadıkları kişilerin çok sayıda görüşünü internet üzerinden okuyabilmektedirler. Genel olarak ifade etmek gerekirse; duygu analizi, sırasıyla amacın tanımlanması ve hedeflerin belirlenmesi, ardından verilerin elde edilmesi, ön işleme, özellik çıkarma, duyarlılığa göre metin sınıflandırması, sonuçların yorumlanması adımlarından oluşan bir süreç olarak görülebilmektedir (Bužić, 2019, s. 216).

Son yıllarda kullanıcılar tarafından sosyal medyada görüş bildirmek amacıyla yoğun olarak kullanılan ve en çok tercih edilen mecranın Twitter olduğu görülmektedir. Bu çalışmada, Twitter ortamında kullanıcılar tarafından atılan tweetler veri kaynağı olarak değerlendirilmiştir. Veri kaynaklarında biriken bu veriler, Twitter API ve farklı yazılımlar aracılığı ile analiz amacıyla çeşitli kodlar kullanılarak çekilmiştir. Söz konusu bu yazılımlardan son yıllarda en popüler olanı Python programlama dili kullanılmıştır. Python; Metin Madenciliği, Derin Öğrenme, Yapay Zeka, vb. konularındaki analizler için oldukça kullanışlı bir programlama dilidir.

Çalışmada kullanılan tweet dili İngilizce'dir. Twitter'dan "5G" ve "vaccination" hashtagleri birlikte kullanılarak elde edilen tweetler ham veri olarak çekilmiş, ön işleme aşamasından geçirilerek temizlendikten sonra geriye kalan tweetler duygu analizi aşamasına dahil edilmiştir. Bu aşamada ham veriler içinde yer alan noktalama işaretleri, semboller, büyüklü küçüklü harfler, sayılar ve linkler ham verilerden ayrıştırılmıştır. Ayrıca zamirler, edatlar, bağlaçlar gibi cümle içinde yer alan; ancak analiz için bir anlamı olmayan kelimeler de "durak kelimeler" (stopwords) olarak adlandırılan liste içine atılarak temizlenmiş ve ön işleme aşaması tamamlanmıştır. Bu kelimelerin analiz edilecek metinden çıkarılmasıyla kelime sayısında düşüş yaşanırken yapılan analizin doğruluk oranı da belirgin bir şekilde artmış olacaktır.

Veri analizi kısmında, ön işleme aşamasına tabi tutularak temizlenen veriler üzerinden "kelime bulutu" (wordcloud) oluşturulmuştur. Bu kelime bulutunun oluşturulabilmesi için Python içinde yer alan numpy ve pandas kütüphanelerinden wordcloud paketi indirilmiştir. Elde edilen kelime bulutu bulgular kısmında Şekil 1'de gösterilmiştir.

Veri analizi aşamasının ardından Duygu Analizi aşamasına geçilmiştir. Duygu Analizi yapabilmek için veriler içerisinde yer alan kelimelerin pozitif, nötr ya da negatif olarak gruplara ayrılması gerekmektedir; dolayısıyla duygu skorları elde edilebilecektir. Bu gruplandırma için öncelikle korpuslar hazırlanmalıdır. Python yazılımında bu korpuslar TextBlob kütüphanesi içerisinde hazır olarak yer almaktadır. Korpuslarda pozitif ve negatif kelimeler bulunmaktadır. Bu korpuslar sayesinde pozitif anlama sahip kelimelerden negatif anlama sahip kelimelerin çıkarılmasıyla duygu skorları elde edilmiş olacaktır. Tweetlerdeki kelimelerden korpustaki pozitif kelime listesi içinde yer alanlar için (+1) puan ve negatif kelime listesi içinde yer alanlar için ise (-1) puan verilmiştir. Bir tweetin 0 puana sahip olması ise; tweetteki duygunun ne pozitif ne de negatif, yani nötr olduğunu göstermektedir. Elde edilen duygu skoru değerleri bulgular kısmında Tablo 2'de gösterilmiştir. Her bir tweet için duygu skorlarının hesaplanmasının ardından elde edilen skorların ortalaması alınarak tweetlerin tamamı için genel skor ortalaması hesaplanmıştır.

Duygu Analizi aşamasının ardından makine öğrenmesi ile 5G ve COVID-19 aşısına yönelik algıya ilişkin yapılan sınıflandırmaların doğruluğunu ölçmek için sınıflandırma algoritmaları kullanılmıştır. Sık kullanılan makine öğrenmesi algoritmalarından CART, NB, KNN ve RF sınıflandırma teknikleri analize dahil edilmiştir. Bunun için veri setinin %70'i eğitim ve %30'u test olacak şekilde iki gruba ayrılmıştır. Algoritmaların kıyaslanabilmesi için model başarım

ölçütleri olan kesinlik, duyarlılık, doğruluk ve F-ölçütü değerleri esas alınmıştır. Her bir algoritma için elde edilen bu değerler bulgular kısmında Tablo 3'te toplu olarak gösterilmiştir. Söz konusu sınıflandırma teknikleri ve model başarım ölçütleri aşağıdaki alt bölümlerde açıklanmıştır.

2.1 Sınıflandırma Teknikleri

Duygu Analizi' nde sınıflandırma için farklı pek çok yaklaşım bulunmaktadır. Bunlar; sözlük tabanlı ve makine öğrenmesi tabanlı yaklaşımlar olarak 2'ye ayrılmaktadır.

Sözlük Tabanlı Yaklaşım: Bu mevcut yaklaşımda, verilen bir metin için mevcut sözlük teknikleri kullanılırken, kelimeler ayrılmaktadır. Genel olarak, puanların bir araya getirilmesiyle skorlama gerçekleştirilir. Örneğin; her bir kelime için ayrı ayrı pozitif, negatif ve nötr gibi öznel kelime puanları toplanır. Her kelimeye bir puan atar ve puan oluşturulur. Maksimum puanı alan, metnin genel bölünmesini verir. Esas olarak sözlüğe dayalı yaklaşım ve korpus tabanlı yaklaşım olmak üzere iki bölümden oluşmaktadır.

Sözlüğe dayalı yaklaşımda eşanlamlılar ve bu sözcüklerin zıttı aranarak ve gruba eklenir. Bu yöntem, sözlük ölçeğine, duyarlılık sınıflandırmasının yoğunluğuna bağlı olarak sakıncalıdır. Sözlük boyutu arttıkça etkisi azalmaktadır. Korpus tabanlı yaklaşımda ise oluşturulan kelimeler içeriğe özeldir ve etiketlenmiş büyük bir veri kümesine ihtiyaç duyulmaktadır.

Makine Öğrenimine Dayalı Yaklaşım: Duyguların sınıflandırılmasında makine öğrenimi teknikleri, metin verilerinde iyi bilinen makine öğrenimi teknolojisinin kullanımına bağlıdır. Duyguların makine öğrenimine göre sınıflandırılması; öncelikli olarak denetimsiz, hibrit tabanlı ve denetimli öğrenme yöntemleri olarak kategorize edilebilir.

Denetimsiz Öğrenme: Bu teknik, denetimli öğrenmenin aksine sınıflandırıcıyı eğitmek için önceden listelenmiş verileri kullanmaz. Denetlenmeyen makine öğrenimi algoritmalarının daha yaygın örnekleri k-means ve apriori algoritmalarıdır.

Hibrit Tabanlı Yaklaşım: Hem makine öğrenimi hem de sözlük tabanlı sınıflandırma yaklaşımını kullanır. Çok az araştırma tekniği, duygu sınıflandırmasını geliştirmek için sözlük tabanlı ve otomatik öğrenme tekniklerinin bir karışımını önerir. Bu hibrit yaklaşım, her ikisinden de en iyisini elde edebileceği için doğruluk oranının artmasından dolayı öncelikli olarak avantajlıdır (Anvar Shathik & Krishna Prasad, 2020, s. 43).

Denetimli öğrenme: Denetimli öğrenme, görüşleri sınıflandırmada etkili bir yöntemidir. Duygu Analizi'nde kullanılan pek çok denetimli sınıflandırma tekniği bulunmaktadır. Aşağıda sadece bu çalışmada kıyaslama amacıyla kullanılan denetimli sınıflandırma teknikleri açıklanmıştır:

Sınıflandırma ve Regresyon Ağacı (CART)*:* Karar ağacı, örnek uzayının özyinelemeli bir bölümü olarak ifade edilen bir sınıflandırıcıdır (Rokach & Maimon, 2005, s. 166). Bir ağaç yapısı biçiminde sınıflandırma veya regresyon modelleri oluşturur. Bir veri kümesini daha küçük alt kümelere ayırırken, aynı zamanda ilişkili bir karar ağacı aşamalı olarak geliştirilir. Nihai sonuç, karar düğümleri ve yaprak düğümleri olan bir ağaçtır. Bir karar düğümünün iki veya daha fazla dalı vardır. Kök düğüm adı verilen en iyi tahmin ediciye karşılık gelen, bir ağaçtaki en üstteki karar düğümüdür. Yaprak düğümü, bir sınıflandırma veya kararı temsil eder. Karar ağaçları hem kategorik hem de sayısal verileri işleyebilir (Sayad, t.y.).

1984 yılında Breiman tarafından ortaya konulan CART algoritması, bir ikili sınıflandırma yöntemidir. Bölme koşulu Gini indeksine göre belirlenmektedir. Her bölme işlemi, verilerin iki alt gruba bölünmesini gerektirmektedir. Daha sonra her alt küme, bir sonraki test özelliğini belirlemek için ayrıca bölünür. Bölme işlemi, veriler artık bölünemeyene kadar devam eder.

D, n adet örnek içeren bir veri seti ve *Pj* de, j kategorisinde yer alan D'nin göreli olasılığı olsun. Gini katsayısı şu şekilde ifade edilir:

$$\mathrm{Gini(D)} = 1 - \sum_{j=1}^{n} P_j^2 \quad (1)$$

Gini katsayısının amacı, verilerdeki en fazla sayıda kategoriyi farklı düğümlerdeki diğer kategorilerden ayırmaktır. Gini değeri daha küçük olduğunda, örneğe ait kategori dağılımı daha düzensizdir. Bu, bölme noktası kullanılarak oluşturulan alt kümedeki kategori saflığının daha yüksek olması durumunda, farklı kategoriler arasında ayrım yapma yeteneğinin artacağı anlamına gelir.

Kass (1980) tarafından önerilen CHAID algoritmasında, bölünme koşulunu belirlemek için kesikli bir test (χ^2, ki-kare istatistiği) uygulanmıştır. Esas olarak birkaç değişken arasındaki bağımlılık derecesini hesaplamak için kullanılır. χ^2 ile hesaplanan değer ne kadar büyükse, değişkenin bağımlılık derecesi ve olasılık değeri o kadar yüksek olur.

QUEST, ağaç yapılarını sınıflandırmak için kullanılan bir algoritmadır ve Loh ve Shih (1997) tarafından önerilmiştir. Bu algoritmadaki bölme kuralı, hedef değişkenin sürekli bir değişken olduğu varsayımını içerir. Hesaplama hızı diğer yöntemlerden daha yüksektir ve QUEST algoritması birden çok kategori değişkeni için daha uygundur, ancak yalnızca ikili verileri işleyebilmektedir.

1983 yılında J. R. Quinlan tarafından önerilen ID3 karar ağaçlarında (Iterative Dichotomizer 3) entropi kavramı yer almaktadır. *Entropi*, bir sistemin düzenlenebileceği rastgele yolların sayısının bir ölçüsüdür. N adet kayıt içeren S veri seti için bilgi entropisi şu şekilde tanımlanmaktadır:

$$\mathrm{Entropi(S)} = -\sum P_1 \log_2 P_1 \quad (2)$$

Burada P_1, sınıf 1'e ait olan S'nin oranıdır.

Gain ise, örneklerin belirli bir özniteliğe göre sınıflandırılmasının neden olduğu bilgi entropisinde beklenen azalmadır. S veri setindeki A özniteliğinin bilgi kazancı şu şekilde tanımlanmaktadır:

$$\mathrm{Gain(S,A)} = \mathrm{Entropi(S)} - \sum_{v \epsilon \mathrm{Value(A)}} \frac{|S_v|}{|S|} \mathrm{Entropi}\,(S_v) \quad (3)$$

Burada A, A özniteliğinin tüm olası değerlerinin kümesidir. S_v, A özniteliği için S'nin alt kümesidir. $|S_v|$, S_v'deki elemanların sayısı ve $|S|$ de S'deki elemanların sayısıdır.

C4.5 (Commercial Version 4.5), 1993 yılında ortaya konulmuş ID3'ün geliştirilmiş bir sürümüdür. ID3'te karar ağacının oluşturulması sırasında sınıflandırma kriteri olarak *gain* kullanılıyorken, C4.5'te *kazanım oranı* kullanılır. Bu algoritma hem sürekli hem de ayrık öznitelikleri işler. Sürekli öznitelikleri işlemek için C4.5 bir eşik oluşturur ve ardından; birincisi özniteliğin değerine sahip grup için eşik değerinin üstünde, ikincisi ise özniteliğin değerine sahip grup için eşik değerine eşit ya da eşik değerinden daha küçük olacak şekilde listeyi iki kategoriye ayırır.

ID3'e benzer şekilde, en iyi sınıflandırma özniteliğini belirlemek için veriler ağacın her düğümünde sıralanır (Almunirawi & Maghari, 2016, s. 754).

Naïve Bayes (NB): Veri kümesindeki farklı değişkenler arasında koşullu bağımsızlık varsayımına dayanan olasılıksal bir sınıflandırma yöntemidir (Go vd., 2009, s. 3). Bayes teoremini esas almaktadır.

Olasılık teorisinde, Bayes teoremi iki rasgele olayın koşullu ve marjinal olasılıklarını ilişkilendirmektedir. Gözlemler genellikle verilen posterior olasılıkları hesaplamak için kullanılmaktadır. Verilen bir $x = (x_1, x_2, \dots, x_n)$ vektörü için k adet olasılık varken olasılık $p(C_k|x_1, x_2, \dots, x_n)$ olmaktadır. Bayes teoremi uygulandığında;

$$\mathrm{p(C_k|x)} = \frac{\mathrm{p(x|C_k)p(C_k)}}{\mathrm{p(x)}} = \frac{\mathrm{p(x_1, x_2, \dots, x_k\,|C_k)p(C_k)}}{\mathrm{p(x_1, x_2, \dots, x_k)}} \tag{4}$$

elde edilir.

Her özelliğin koşullu olarak diğer tüm özelliklerden bağımsız olduğunu varsaydığımızda, tüm sınıflar için payda sabit kaldığından dolayı sadece pay kullanılmalıdır. Böylelikle Naïve Bayes olasılık modeli elde edilmiş olur (Singh, 2015, s. 3).

$$\begin{aligned}\mathrm{p(C_k|x)} &\propto \mathrm{p(C_k)p(x_1|C_k)p(x_2|C_k)\dots p(x_n|C_k)} \\ &\propto \mathrm{p(C_k)}\prod_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{p(x_i|C_k)}\end{aligned} \tag{5}$$

Naïve Bayes Sınıflandırıcı, $p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$ için maksimum değere sahip sınıfı seçmektedir.

k-En Yakın Komşuluk (KNN): Uzaklık hesaplamasına dayalı olarak en yakın k verisini bulmak için eğitim veri kümesi ve önceden tanımlanmış bir k değeri gerektiren bir öğrenme algoritmasıdır. k verisinin farklı sınıfları varsa, algoritma bilinmeyen verilerin sınıfının çoğunluk sınıfıyla aynı olacağını tahmin eder. KNN'de en yakın komşuluk, uzaklık ölçülerinin hesaplanmasıyla bulunmaktadır. Uzaklık ölçüleri, yeni bir veri noktası ile mevcut eğitim veri kümesi arasındaki mesafeyi bulmayı sağlamaktadır.

Öklid Uzaklığı: İki nokta arasındaki uzaklığı bulmak için kullanılmaktadır. Sınıflandırma için en çok kullanılan ölçüdür (Chomboon vd., 2015, s. 281). Ölçülmek istenen uzaklık $X = (x_1, x_2, \dots, x_n)$ ve $Y = (y_1, y_2, \dots, y_n)$ olarak ele alınsın. Bu durumda *X* ve *Y* noktaları arasındaki Öklid uzaklığını hesaplamak için aşağıdaki bağıntı kullanılır:

$$\mathrm{D(X,Y)} = \sqrt{\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i} - \mathrm{y_i})^2} \tag{6}$$

Manhattan Uzaklığı: *X* ve *Y* noktaları arasındaki uzaklık, kartezyen koordinatların mutlak farkının toplamıdır. *City-blok uzaklığı* olarak da bilinmektedir.

$$\mathrm{D(X,Y)} = \sum_{\mathrm{i=1}}^{\mathrm{n}}|\mathrm{x_i} - \mathrm{y_i}| \tag{7}$$

Minkowski Uzaklığı: Öklid ve Manhattan uzaklıklarının genelleştirilmiş halidir. Aşağıdaki bağıntıdan hareketle hesaplanır:

$$\mathrm{D(X,Y)} = \left[\sum_{\mathrm{i=1}}^{\mathrm{n}}|\mathrm{x_i} - \mathrm{y_i}|^{\mathrm{m}}\right]^{1/\mathrm{m}} \tag{8}$$

Minkowski uzaklığının özel durumları bulunmaktadır:

$m = 1$ iken; Minkowski uzaklığı, Manhattan uzaklığını verir,

$m = 2$ iken; Minkowski uzaklığı, Öklid uzaklığını verir,

$m = \infty$ iken; Minkowski uzaklığı, Chebyshev uzaklığını verir.

Chebyshev Uzaklığı: İki vektör veya standart koordinatlı noktalar arasındaki uzaklığı bulmak için kullanılan bir ölçüdür. Aşağıdaki bağıntı ile hesaplanmaktadır (Oğuzlar & Kızılkaya, 2019, s. 48):

$$\lim_{m\to\infty}\left[\sum_{i=1}^{n}|x_i-y_i|^m\right]^{1/m} \quad (9)$$

Rastgele Orman (RF): Rastgele seçilen eğitim kümesi alt kümesinden bir dizi karar ağacı oluşturmaktadır. Ardından, test veri setinin son sınıfına karar vermek için farklı karar ağaçlarından alınan değerleri toplamaktadır (Patel, 2017). Rastgele Orman sınıflandırıcısı aynı zamanda, hem regresyon hem de (çok sınıflı) sınıflandırmayı ele almaktadır. Eğitilmesi ve tahmin edilmesi nispeten hızlıdır. Yalnızca bir veya iki ayar parametresine bağlıdır. Yerleşik bir genelleme hatası tahminine sahiptir ve doğrudan yüksek boyut için kullanılabilmektedir (Cutler, vd., 2012, s. 1).

2.2 Model Başarım Ölçütleri

Denetimli Makine Öğrenmesinde, öğrenme algoritmalarının ve sınıflandırıcıların performansını değerlendirmenin farklı yolları vardır. Model Başarım ölçüleri, her sınıf için doğru ve yanlış şekilde tanınan örneklerin kaydedildiği bir karışıklık matrisinden oluşturulur (Sokolova, vd., 2006, s. 1). Karışıklık matrisi, sınıflandırma problemlerini çözerken kullanılan çok popüler bir ölçüdür. Bu matris, ikili sınıflandırma ve çok sınıflı sınıflandırma problemlerinde uygulanabilmektedir (Kulkarni vd., 2020, s. 83). Tablo 1'de, ikili sınıflandırma için bir karışıklık matrisi verilmiştir (Sokolova, vd., 2006, s. 1).

Tablo 1

Karışıklık matrisi

		Tahmin Edilen	
		Pozitif	Negatif
Gerçek	Pozitif	Doğru Pozitif (TP)	Yanlış Negatif (FN)
	Negatif	Yanlış Pozitif (FP)	Doğru Negatif (TN)

Sınıflandırma için kullanılan nxn boyutundaki bir karışıklık matrisi, tahmin edilen ve gerçek sınıflandırmayı göstermektedir. Burada n, farklı sınıfların sayısıdır. Tablo 1, $n = 2$ için bir karışıklık matrisini göstermektedir (Visa vd., 2011, s. 3).

Verilen karışıklık matrisinde;

TP: Pozitif tahminler içerisindeki doğru sınıflandırılmış tahmin sayısını,
FP: Pozitif tahminler içerisindeki yanlış sınıflandırılmış tahmin sayısını,
TN: Negatif tahminler içerisindeki doğru sınıflandırılmış tahmin sayısını,
FN: Negatif tahminler içerisindeki yanlış sınıflandırılmış tahmin sayısını

ifade etmektedir. Bu tanımlamalar üzerinden elde edilebilen sınıflandırma ölçütleri şunlardır:

Doğruluk (Accuracy - A): Doğru sınıflandırılmış örnek sayısının toplam örnek sayısına oranlanmasıyla elde edilir. Model başarım ölçütleri arasında en sık kullanılan ölçüttür. Bu ölçüt,

$$A=\frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

oranı ile hesaplanır. Modelin hatası ise,

$$\text{Hata} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (11)$$

oranı ile hesaplanır. Aynı şekilde bu değer, doğruluk ölçütü üzerinden de hesaplanabilir.

$$\text{Hata} = 1 - \text{A} \quad (12)$$

Kesinlik (Precision - P): Doğru sınıflandırılmış pozitif örnek sayısının toplam pozitif tahmin edilen örnek sayısına oranıdır.

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (13)$$

Duyarlılık (Recall - R): Doğru sınıflandırılmış pozitif örnek sayısının, toplam pozitif örnek sayısına oranıdır. *Kesinlik* ile *duyarlılık* arasında ters orantı bulunmaktadır.

$$\text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (14)$$

F-ölçütü (F): Model başarım ölçütlerinden *kesinlik* ile *duyarlılık* ölçütlerinin harmonik ortalamasıdır.

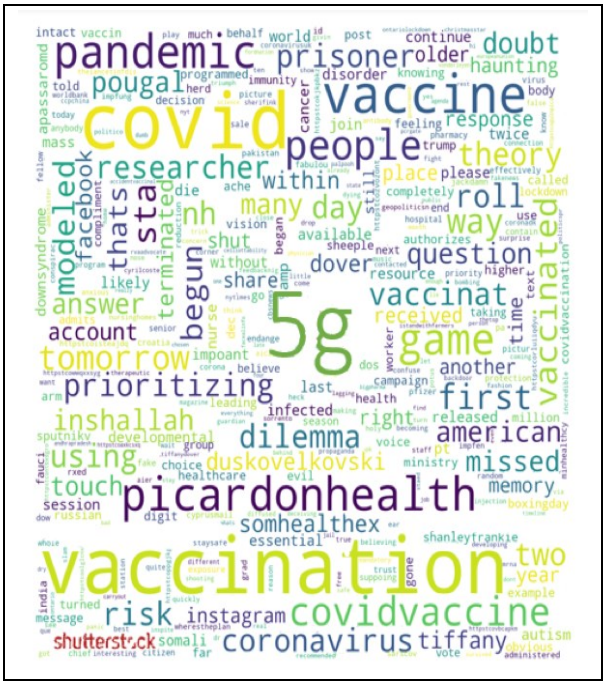
$$\text{F} = 2 * \frac{\text{P} * \text{R}}{\text{P} + \text{R}} \quad (15)$$

3. Bulgular

Twitter'dan "5G" ve "vaccination" hashtagleri kullanılarak toplam 25642 adet tweet (ham veri) çekilmiş, ön işleme aşamasından geçirildikten sonra temizlenen verilerin 19634 adet olduğu gözlemlenmiştir. Veri analizi kısmında, temizlenen veriler üzerinden elde edilen "kelime bulutu" (wordcloud) Şekil 1'de görülmektedir.

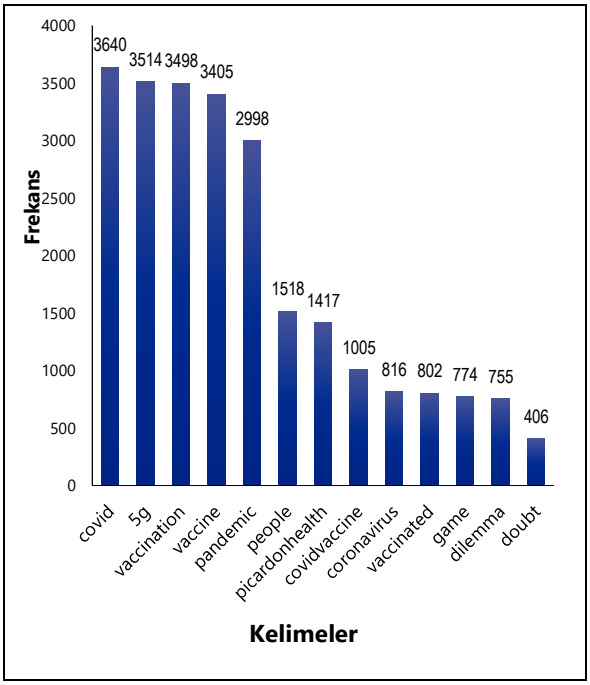
Şekil 1

Kelime Bulutu



Şekil 2

Kelimelere Ait Frekans Grafiği



Görselde yer alan kelimeler, metin içerisinde bulunan kelimelerin frekanslarına göre oluşturulmuştur. Elde edilen kelime bulutuna göre, puntosu büyük olan kelimelerin frekansları da en yüksektir. Buna göre; *covid*, *5g*, *vaccine, vaccine* ve *pandemic* kelimelerinin metin içinde en sık gözlenen kelimeler olduğu gözlemlenmektedir. Benzer şekilde *people*, *picardonhealth*, *covidvaccine*, *coronavirus*, *vaccinated*, *game*, *dilemma* ve *doubt* kelimelerinin frekansları da; frekansları en yüksek ilk 5 kelimeye oranla daha düşük, ancak metinde yer alan diğer kelimelere oranla çok daha yüksektir. Bu kelimeler ve sahip oldukları yüksek frekansları araştırma konusuyla uyumluluk sağlamaktadır. Metinde yer alan frekansı en yüksek kelimelere ilişkin çubuk grafiği Şekil 2'de gösterilmiştir.

Veri analizi aşamasının ardından elde edilen duygu skoru değerleri Tablo 2'de gösterilmiştir.

Tablo 2

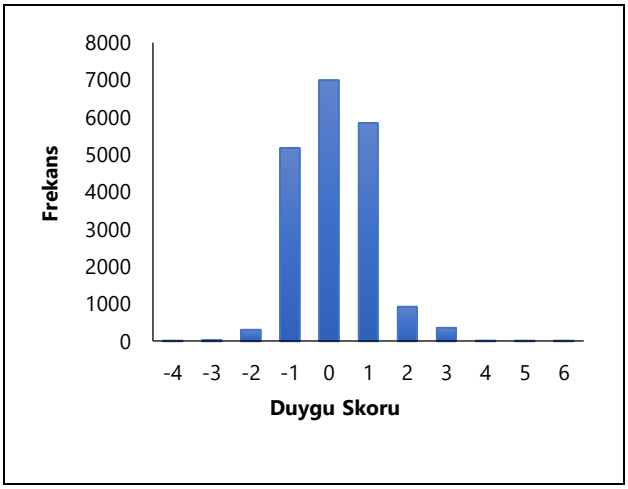
Duygu Skoru Değerleri

-4	-3	-2	-1	0	1	2	3	4	5	6
8	19	294	5176	6990	5852	916	358	16	4	1

Tablo 2'de yer alan tüm duygu skoru değerleri -4 ile 6 arasında değişmektedir. Bu skorlar, tweetlerdeki pozitif kelimelerin toplamı ile negatif kelimelerin toplamı arasındaki farktan elde edilmiştir. Elde edilen duygu skoru değerlerine bakıldığında; analizde yer alan tweetler içerisinde duygu skoru (-4) olan 8 adet tweet, (-3) olan 19 adet tweet, (-2) olan 294 adet tweet ve (-1) olan 5176 adet tweet olduğu görülmektedir. Duygu skoru değeri 0 olarak elde edilen tweet sayısı ise 6990'dır. Ayrıca analizde yer alan tweetler içerisinde duygu skoru 1 olan 5852 adet tweet, 2 olan 916 adet tweet, 3 olan 358 adet tweet, 4 olan 16 adet tweet, 5 olan 4 adet tweet ve 6 olan 1 adet tweet bulunmaktadır. Pozitif tweetlerin sayısı, negatif tweetlerin sayısının 1,3 katıdır. Tweetlerin tamamı için duygu skoru değerlerinin hesaplanmasının ardından elde edilen genel duygu skoru ortalaması ise 0,15 olarak bulunmuştur. Elde edilen duygu skoru grafiği ise Şekil 3'te gösterilmiştir.

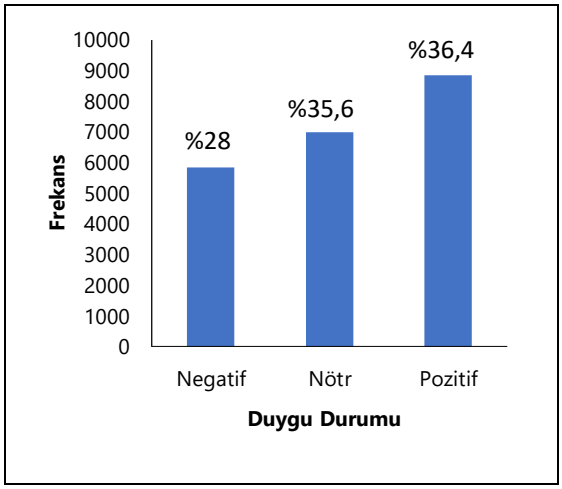
Şekil 3

Duygu Skoru Grafiği



Şekil 4

Duygu Durumu Grafiği



Negatif, nötr ve pozitif duygu skoruna sahip tweetlerin toplam tweetler içindeki yüzdelik oranları Şekil 4'teki grafikte gösterilmiştir.

Elde edilen duygu skoru değerlerine göre, dünya genelinde İngilizce tweet atan kişilerin 5G ve COVID-19 aşısı hakkında bildirdikleri görüşler hakkında çok ciddi bir farklılık olmadığı görülmektedir. Elde edilen sonuçlar incelendiğinde, pozitif görüş bildiren kişilerin attığı tweet sayısı 8848'dir ve bu tweetlerin oranı %36,4'tür. Bu konu hakkında bir memnuniyet ya da rahatsızlık duymayan kişilerin sayısı da 6990 olarak elde edilmiştir ve %35,6'lık bir oran ile pozitif fikir beyan eden kişi sayısıyla oldukça yakın olduğu görülmektedir. 5G teknolojisinin ve COVID-19 aşısının negatif etkilerinin olduğunu düşünen kişilerin sayısı ise 5853 ve tweet oranı ise %28'dir.

Analize dahil edilen makine öğrenmesi algoritmalarından CART, NB, KNN ve RF algoritmalarının kıyaslanabilmesi için veri seti %70'i eğitim ve %30'u test olacak şekilde iki gruba ayrıldıktan sonra her bir algoritma için elde edilen kesinlik (P), duyarlılık (R), doğruluk (A) ve F-ölçütü (F) değerleri Tablo 3'te toplu olarak gösterilmiştir.

Tablo 3

Makine Öğrenmesi Sonuçları

Algoritma	P	R	A	F
CART	0,5308	0,8178	0,6926	0,6125
NB	**0,7852**	0,4643	**0,7445**	0,6234
KNN	0,6127	**0,8209**	0,6883	0,5644
RF	0,4428	0,6892	0,7402	**0,7866**

Makine öğrenmesi sonuçlarına göre, 5G ile COVID-19 aşısı arasında var olabileceği söz konusu olan ilişkiyi analiz etmek adına yapılan sınıflandırmada algoritmalar farklı sınıflandırma ölçütlerine göre birbirlerine üstünlük göstermişlerdir. Buna göre yapılan analizde kesinlik (P) ölçütü için en iyi sonucu 0,7852 değeri ile NB ve duyarlılık (R) ölçütü için en iyi sonucu 0,8209 değeri ile KNN algoritmaları vermiştir. Doğruluk (A) ölçütü için ise en iyi sonucu 0,7445 değeri ile yine NB algoritması ve F-ölçütü (F) için en iyi sonucu da 0,7866 değeri ile RF algoritması vermiştir.

Sonuç olarak, elde edilen sonuçlara göre ölçütlerin performansları kıyaslandığında kesinlik (P) ve doğruluk (A) ölçütleri için en iyi sonucu NB, duyarlılık (R) ölçütü için en iyi sonucu KNN ve F-ölçütü (F) için en iyi sonucu RF algoritmasının verdiği görülmektedir.

4. Sonuç

Çalışmada, son zamanlarda sıkça tartışma konusu olarak gündemde yer alan 5G teknolojisi ile COVID-19 aşısı arasındaki olası bir ilişkinin dünya genelinde insanlar üzerinde bir etkiye neden olup olmadığı, yani insanların konuya ilişkin algısı ele alınmıştır.

Söz konusu tartışmaların ne yönde seyrettiğini belirlemek ve insanların genel görüşünü ortaya koyabilmek için son yıllarda oldukça rağbet gören ve insanların rahatça görüşlerini tweet atarak dile getirebildikleri bir sosyal medya ortamı olan Twitter'da paylaşılan yorumlar üzerinden Duygu Analizi yapılmıştır. Bunun için Twitter'dan Twitter API aracılığı ile Ekim – Aralık 2020 tarihleri arasında "5G" ve "*vaccination*" hashtagleri kullanılarak 25642 adet tweet çekilmiştir. Veri analizi aşamasında verilerin temizlenmesi ile birlikte tweet sayısı 19634'a düşmüştür.

Veri analizi aşamasında çekilen tweetlerde yer alan kelimelerin frekansları elde edilmiştir. Buna göre, tweetlerde en sık geçen kelimelerin sırasıyla c*ovid*, *5g*, *vaccine*, *vaccine* ve *pandemic* olduğu görülmektedir. Benzer şekilde *people*, *picardonhealth*, *covidvaccine*, *coronavirus*,

vaccinated, *game*, *dilemma* ve *doubt* kelimelerinin frekanslarının da belirtilen kelimelere kıyasla daha düşük frekansa sahip olmasına rağmen diğer kelimelere göre nispeten daha yüksek frekansa sahip oldukları görülmüştür.

Duygu analizi aşamasında ise, tweetlerde yer alan kelimelerdeki duygunun ortaya çıkarılabilmesi amacıyla kelimelere ilişkin duygu skorları belirlenmiştir. Bu amaçla kelimeler pozitif, nötr ve negatif olmak üzere gruplara ayrılmıştır. Elde edilen duygu skoru değerlerine göre, dünya genelinde İngilizce tweet atan kişilerden %35,6'sının 5G ve COVID-19 aşısı hakkında genel olarak bir memnuniyet ya da rahatsızlık duymadıkları söylenebilir. Tweet atanların yaklaşık olarak %36,4'ünün konu hakkında memnuniyet duyduğu ortaya çıkmıştır. Geriye kalan %28'lik kesimin ise, 5G ve COVID-19 aşısı arasında bulunabilecek bir ilişkiden rahatsızlık duyduğu sonucuna varılmıştır. Genel duygu skoru ortalaması ise 0,15 olarak bulunmuştur.

Makine öğrenmesi aşamasında bulunan değerlere CART, NB, KNN ve RF algoritmaları uygulanmıştır. Bu amaçla veri seti %70 eğitim ve %30 test seti olarak iki gruba ayrılmıştır. Elde edilen sonuçlara göre, NB algoritmasının kesinlik (P) ve doğruluk (A) ölçütleri için en iyi sonuçları verdiği görülmektedir. Bu durum sırası ile 0,7852 ve 0,7445 değerleri ile sağlanmıştır. KNN algoritmasının ise duyarlılık (R) ölçütü için en iyi performansı sağladığı görülmektedir. Bu ölçütün değeri ise 0,8209 olarak elde edilmiştir. RF algoritmasının en iyi performansı da F-ölçütü (F) için 0,7866 değeri ile gösterdiği görülmektedir.

Elde edilen sonuçlara göre; 5G ile COVID-19 aşısı arasında var olabileceği söz konusu olan ilişkiyi analiz etmek adına yapılan sınıflandırmada en iyi sonuçları kesinlik (P) ve doğruluk (A) ölçütleri için NB, duyarlılık (R) ölçütü için KNN ve F-ölçütü (F) için de RF algoritmaları vermiştir.

FİNANSAL DESTEK
Yazarlar bu çalışma için herhangi bir finansal destek almadıklarını beyan etmişlerdir.

ETİK
Yazarlar bu çalışmada etik ilke ve standartlara uyduklarını beyan etmişlerdir.

YAZAR KATKI BEYANI
Elçin Timur Çakmak | Kavram/fikir; Literatür taraması; Tasarım; Veri toplama/analiz; Veri/bulguların yorumu; Taslağın yazımı; Son onay ve sorumluluk. Genel katkı düzeyi %65

Ayşe Oğuzlar | Tasarım; Yönetme ve kontrol; Eleştirel inceleme; Son onay ve sorumluluk. Genel katkı düzeyi %35.

ÇIKAR ÇATIŞMASI
Yazarlar herhangi bir çıkar çatışması beyan etmemiştir.

Kaynakça

Adwan, O. Y., Al-Tawil, M., Huneiti, A., Shahin, R., Zayed, A. A., & Al-Dibsi, R. (2020). Twitter sentiment analysis approaches: A survey. *International Journal of Emerging Technologies in Learning (IJET)*, *15*(15), 79–93. https://doi.org/10.3991/ijet.v15i15.14467

Almunirawi, K. M., & Maghari, A. Y. A. (2016). A comparative study on serial decision tree classification algorithms in text mining. *International Journal of Intelligent Computing Research (IJICR)*, 7(4), 754–760.

Anvar Shathik, J. & Krishna Prasad, K. (2020). A literature review on application of sentiment analysis using machine learning techniques. *International Journal of Applied Engineering and Management Letters (IJAEML)*, *4*(2), 41–77. http://doi.org/10.5281/zenodo.3977576

Bahja, M., & Safdar, G. A. (2020). Unlink the link between COVID-19 and 5G networks: An NLP and SNA based approach. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 209127-209137. https://doi.org/10.1109/ACCESS.2020.3039168

Bužić, D. (2019). Sentiment analysis of text documents. In V. Strahonja & V. Kirinić (Eds.), Proceedings of the Central European conference on information and intelligent systems (pp. 215–221). University of Zagreb.

Chomboon, K., Chujai, P., Teerarassamee, P., Kerdprasop, K., Kerdprasop, N. (2015). An empirical study of distance metrics for k-nearest neighbor algorithm. In *Proceedings of the 3rd international conference on industrial application engineering* (pp. 280–285). The Institute of Industrial Applications Engineers, Japan. Doi: https://doi.org/10.12792/iciae2015.051

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. In C. Zhang & Y. Ma (Ed.), *Ensemble Machine Learning: Methods and Applications* (pp. 157–175). Springer US. https://doi.org/10.1007/978-1-4419-9326-7_5

Go, A., Huang, L., & Bhayani, R. (2009, June 6). *Twitter sentiment analysis* [Final project report CS224N]. The Stanford NLP Group. https://www-nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf

Ha, H., Han, H., Mun, S., Bae, S., Lee, J., & Lee, K. (2019). An improved study of multilevel semantic network visualization for analyzing sentiment word of movie review data. *Applied Sciences*, *9*(12), 8–31. https://doi.org/10.3390/app9122419

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2), 119–127. https://doi.org/10.2307/2986296

Kim, H., & Jeong, Y.-S. (2019). Sentiment classification using convolutional neural networks. *Applied Sciences*, *9*(11), 2347. https://doi.org/10.3390/app9112347

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). 5—Foundations of data imbalance and solutions for a data democracy. In F. A. Batarseh & R. Yang (Ed.), *Data democracy: At the nexus of artificial intelligence, software development, and knowledge engineering* (pp. 83–106). Academic Press. https://doi.org/10.1016/B978-0-12-818366-3.00005-8

Kumar, T.S.S., Devi, N.T.D., Krishnendhu, T.K., Neethu, K.E., Radhakrishnan, S. C. (2020). Review of sentiment analysis: A multilingual approach. *International Journal of Advanced Research in Computer and Communication Engineering*, *9*(1), 53–58.

Loh, W.Y., & Shih, Y.S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*(4), 815–840.

Mehta, P., & Pandya, S. (2020). A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific & Technology Research*, *9*(2), 601–609. http://www.ijstr.org/final-print/feb2020/A-Review-On-Sentiment-Analysis-Methodologies-Practices-And-Applications.pdf

Oğuzlar, A., & Kızılkaya, Y. M. (2019). *Metin madenciliğinde duygu analizi - R uygulamalı*. Dora Yayınevi.

Patel, S. (2017, May 18). Chapter 5: Random Forest Classifier. *Machine Learning 101*. https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1

Piksina, O., & Vernholmen, P. (2020). Coronavirus-Related Sentiment and Stock Prices Measuring Sentiment Effects on Swedish Stock Indices (Degree Project). Real Estate and Finance, Institutionen För Fastigheter Och Byggande, Stockholm, Sweden. https://www.diva-portal.org/smash/get/diva2:1442317/FULLTEXT01.pdf

Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). *Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research* (arXiv:2005.00357). arXiv. https://doi.org/10.48550/arXiv.2005.00357

Qualcomm (t.y.). Everything you need to know about 5G. Qualcomm Technologies, Inc. Retrieved March 25, 2021, from https://www.qualcomm.com/5g/what-is-5g

Rajput, N. K., Grover, B. A., & Rathi, V. K. (2020). *Word frequency and sentiment analysis of Twitter messages during coronavirus pandemic* (arXiv:2004.03925). arXiv. https://doi.org/10.48550/arXiv.2004.03925

Reality Check Team. (2019, July 15). *Does 5G pose health risks?* BBC News. https://www.bbc.com/news/world-europe-48616174

Rokach, L., & Maimon, O. (2005). Decision trees. In O. Maimon & L. Rokach (Ed.), *Data mining and knowledge discovery handbook* (pp. 165–192). Springer US. https://doi.org/10.1007/0-387-25465-X_9

Sayad, S. (t.y.). *Decision tree-classification*. An introduction to data science. Retriewed April 3, 2021, from https://www.saedsayad.com/decision_tree.htm

Singh, A. (2015). *Twitter sentiment analysis* (Report No. CS365A: 12056). https://cse.iitk.ac.in/users/cs365/2015/_submissions/ajaysi/report.pdf

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *Proceedings of the 19th Australian joint conference on Artificial Intelligence: Advances in artificial intelligence*, 1015–1021. https://doi.org/10.1007/11941439_114

Tian, L., Lai, C., & Moore, J. (2018). Polarity and intensity: The two aspects of sentiment analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)* (pp. 40–47). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/W18-3306

Timur Çakmak, E., & Oğuzlar, A. (2020). 2020 ABD başkanlık seçimleri üzerine sosyal medya duygu analizi. İçinde *20. Uluslararası ekonometri, yöneylem araştırması ve istatistik sempozyumu tam metin kitapçığı* (ss. 19–27). Ankara Hacı Bayram Veli Üniversitesi.

Visa, S., Ramsay, B., Ralescu, A., & van der Knaap, E. (2011). Confusion matrix-based feature selection. In S. Visa, A. Inoue, & A. Ralescu (Ed.), *Proceedings of the 22nd Midwest Artificial Intelligence and Cognitive Science Conference* (pp. 120–127). Cincinnati.

Wilson, S. L., & Wiysonge, C. (2020). Social media and vaccine hesitancy. *BMJ Global Health*, *5*(10), e004206. https://doi.org/10.1136/bmjgh-2020-004206

Extended Abstract

In this study; 5G technology, which has been frequently handled as a topic of discussion, is examined. While it is already a subject open to criticism, with the COVID-19 vaccine came into our lives, there have been discussions around the world about whether there is a relationship between them. In order to determine the direction of the aforementioned discussions and to reveal the general opinion of the people, Sentiment Analysis was carried out through the tweets shared on Twitter, a social media environment where people can easily express their opinions by tweeting. To this end, 25642 tweets were drawn using the "5G" and "vaccination" hashtags between October and December 2020 via Twitter API. During the data analysis phase, the number of tweets decreased to 19634 with the cleaning of the data.

During the data analysis phase, the frequencies of the words in the tweets were obtained. Accordingly, it is seen that the most common words in tweets are *covid*, *5g*, *vaccine*, *vaccine* and *pandemic*, respectively. Similarly, although the frequencies of the words *people*, *picardonhealth*, *covidvaccine*, *coronavirus*, *vaccinated*, *game*, *dilemma* and *doubt* have a lower frequency compared to the mentioned words, they have a relatively higher frequency compared to other words.

In the sentiment analysis stage, sentiment scores for the words were determined in order to reveal the emotion in the words in the tweets. For this purpose, the words were divided into groups as positive, neutral and negative. According to the emotion score values obtained, it can be said that 35.6% of the people who tweet in English around the world are not generally satisfied or uncomfortable about 5G and the COVID-19 vaccine. It has been revealed that approximately 36.4% of tweeters are satisfied with the subject. It was concluded that the remaining 28% were uncomfortable with the relationship that could be found between 5G and the COVID-19 vaccine, also the overall sentiment score is 0.15.

In the machine learning stage; CART, NB, KNN ve RF algorithms were applied to the values to find the possible relationship between 5G and the COVID-19 vaccine in the classification. For this purpose, the data set was divided into 2 groups as 70% training and 30% test set. According to the analysis; the best results were obtained by NB with 0.7852 precision (P) and 0.7445 accuracy (A) values, KNN with 0.8209 recall (R) value, and RF with 0.7866 F-measure (F) value.