Türk Kütüphaneciliği, 36, 4 (2022), 472-494
Doi:10.24146/1152238

Bilgi Yönetimi Bağlamında Metin Madenciliği Teknikleri ile Dijital İçerik Analizi

Digital Content Analysis with Text Mining Techniques in the Context of Information Management

Levent Kurt [*], **Oya Gürdal** [**] ***ve İnci Batmaz*** [***]

Öz

Amaç: *Bu çalışmada, blockchain teknolojileri konusunda internet üzerinde içerik yayınlayan bir platformun içerik analizi yapılmıştır. Araştırmanın amacı, platformun Facebook'ta paylaştığı içerikler için başlık bazında okunma oranını etkileyen faktörlerin (*kelime ve kelime gruplarının*) tespit edilmesidir.*

Yöntem: *Araştırma sınırlılıkları kapsamında belirlenen tarih aralığında yayınlanan 2206 içerikten 500 tanesi rastgele seçilmiştir. İçeriklerin başlıkları* Python programlama dili *kullanılarak bu çalışmadaki probleme özel olarak farklı bir yaklaşımla ve standart metin madenciliği teknikleriyle çözümlenmiş ve metinler üzerinden yapısallaştırılmış iki farklı veri kümesi elde edilmiştir. Elde edilen iki farklı veri kümesi üzerinde çoklu doğrusal regresyon yöntemi kullanılarak analizler gerçekleştirilmiştir.*

Bulgular: *Analizler sonucunda içerik başlıklarında kullanılan bazı kelime ve kelime gruplarının, içeriklerin okunma oranını etkilediği tespit edilmiştir. Ayrıca uygulanan farklı yaklaşımın standart metin madenciliği tekniklerine göre daha yüksek performans sağladığı belirlenmiştir.*

Sonuç: *Araştırmada ham veri işlenerek değerli bilgiler elde edilmiştir. Teorik olarak ortaya çıkarılan bilgiler, uygulama pratiğiyle karşılaştırılmış ve tutarlı sonuçlar elde edildiği gözlemlenmiştir. Uygulanan farklı yaklaşımın etkili bir şekilde benzer metin madenciliği problemlerinde kullanılabileceği saptanmıştır.*

Özgünlük: *Araştırmada içerik başlığı bazında yapılan metin madenciliğine dayalı analiz, farklı bir yaklaşımla ele alınmıştır. Bu yönüyle çalışma özgün bir nitelik taşımaktadır.*

Anahtar Sözcükler: *Metin madenciliği; veri madenciliği; çoklu doğrusal regresyon; içerik analizi; bilgi yönetimi.*

[*] Ankara Üniversitesi, Dil ve Tarih-Coğrafya Fakültesi, Bilgi ve Belge Yönetimi Bölümü. E- posta: lewentkurt@gmail.com
Ankara University, Language and History-Geography Department of Information and Records Management. E-mail: lewentkurt@gmail.com
[**] Ankara Üniversitesi, Dil ve Tarih-Coğrafya Fakültesi, Bilgi ve Belge Yönetimi Bölümü E- posta: ogurdal@ankara.edu.tr
Ankara University, Language and History-Geography Department of Information and Records Management. E-mail: ogurdal@ankara.edu.tr
[***] Orta Doğu Teknik Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü. E-posta: ibatmaz@metu.edu.tr
Middle East Tehcnical University, Faculty of Arts and Sciences, Department of Statistics. E-mail: ibatmaz@metu.edu.tr

Geliş Tarihi - *Received:* 01.08.2022; Kabul Tarihi - *Accepted*: 27.12.2022; Yayımlanma Tarihi - *Published*: 30.12.2022
Editör: Kasım Binici

Abstract

Purpose: *In this study, a content analysis of a platform that publishes content on the internet about blockchain technologies was made. The study aims to determine the factors (word and word string) affecting the reading rate of the digital content -on a titles basis- posted by the platform on Facebook.*

Method: *500 out of 2206 examples of content published between the specified dates were chosen randomly. The titles of the content were processed using standard text mining techniques and a new approach specific to the problem in this study on python programming language and then two different datasets were collected. The datasets were analysed using multiple linear regression.*

Findings: *As a result of the analysis, it was discovered that some words and phrases used in the content titles affected the reading rate of the content. In addition, it has been determined that the new approach provides higher performance than standard text mining techniques.*

Implications: *In this study, valuable information was obtained by processing raw data. As a result of the study, the theory was compared with the practice, and it was observed consistent results. It is determined that the new approach can be used effectively in similar text mining problems.*

Originality: *The research relying on text mining was handled with a new approach on the basis of the content title. In this respect, the study has a unique quality.*

Keywords: *Text mining; data mining; multiple linear regression; content analysis; information management.*

Giriş

Veri madenciliği (VM) ve metin madenciliği (MM) gibi ham veriden bilgiye ulaşmadaki tüm süreçleri ifade eden kavramlar, uzun yıllardır bilgi yönetimi (BY) alanında kendine oldukça fazla yer bulmaktadır. Bunun nedeni, BY sürecinin en önemli aşamalarının, organizasyon içerisinde var olan bilginin ortaya çıkarılması; ham veriden yeni bilgi yaratarak organizasyona katma değer sağlayacak bilginin elde edilmesi ve elde edilen bilginin karar destek sistemlerinde kullanılarak yeni süreçlerin oluşturulması adımlarını içermesi olarak ifade edilebilir. Organizasyon içerisinde yer alan ham verilerin işlenerek yeni bilgilerin ortaya çıkarılması sürecinin tamamı VM kavramı çerçevesinde değerlendirilebilir.

Enformasyon çağında bilgi, rekabet avantajı sağlayan ve girişimleri var eden ve üst düzeye çıkaran hayati derecede önemli, kritik bir organizasyonel kaynak haline gelmiştir. Pek çok organizasyon, büyük miktardaki veriyi toplamakta ve depolamaktadır (Berson, Smith ve Thearling, 1999, s. 29). VM ve MM teknikleri organizasyonlara, pazardaki durumları, ürünleri, müşterileri ve rakip firmaları ile ilgili toplanan tüm bu veriyi yöneterek katma değer yaratan bilgiyi elde etmeleri için çok önemli fırsatlar sunmaktadır.

VM olgusunun çapı, her gün biraz daha genişlemektedir, çünkü bu olgu, organizasyonlara, sahip oldukları veri tabanlarından yararlı modelleri ve eğilimleri ortaya çıkartmak için olanaklar sunmaktadır. Organizasyonlar, iç ve dış sistemlerinde yüklü olan dijital içeriğe erişmek ve depolamak için milyonlarca dolar harcamaktadırlar; ancak VM

uygulamaları yapılmadığında, veri depolarının derinlerinde bulunan gizli/saklı, değerli ve uygulanabilir, yani iş süreçlerine aktarılabilir yararlı bilgiyi ortaya çıkarma avantajını yitirirler. Bu sebeplerle, VM pratiği dünyada giderek daha yaygın hale gelmektedir. VM uygulayan organizasyonlar, iş süreçlerinde kullandıkları değer yaratan bilgiden dolayı rekabet üstünlüğü taşımaktadır (Larose ve Larose, 2014, s. xi).

Bilgi kaynaklarını ve içeriğindeki bilgiyi yönetme, üstesinden gelinmesi için büyük çaba isteyen bir iş sürecidir ve uzmanlık alanıdır. Pek çok organizasyon, sistem içinde bilginin yaratımı, paylaşımı, sistemdeki diğer bilgiler ile bütünleşmesi ve dağıtımı sürecini olanaklı kılmak ve BY uygulamaları ve bu uygulamaları destekleyen enformasyon teknolojilerinin kullanımı için uzman insan gücü istihdam etmektedir (Silwattananusarn ve Tuamsuk, 2012).

BY bir veri kullanımı sürecidir. VM'nin temeli ise veriden kullanılabilir yararlı bilgiyi çekip almak için uygun araçların kullanım sürecini ifade etmektedir (Dawei, 2011). Wang ve Wang (2008), VM'nin BY için iki temel biçimde yararlı olabileceğine dikkat çekmektedir: a) Veri madencileri arasında iş süreçlerine ilişkin genel iş aklının ya da bilgisinin paylaşımı; b) İnsan bilgisinin daha geniş kitlelere yayılması için bir araç olarak VM'ni kullanma. Böylece, VM araçları çok büyük miktardaki veri kümesi içerisinde saklı bilgiyi keşfetmek üzere organizasyona yardımcı olabilmektedir.

Bu bağlamda, bilgi çağına uygun olarak BY kavramındaki önemle vurgulanması gereken konu; sadece bilginin, belgenin veya verinin depolanması ve istenildiğinde kullanıma sunulması değil, aynı zamanda oluşan ham verilerin çözümlenmesi, işlenmesi, analiz edilerek yorumlanması ve son olarak kullanıma sunulmasıdır (Özdemirci, 2018). Hatta son yıllarda yaşanan teknolojik ve bilimsel gelişmelerle birlikte, bazı sektörlerde veri, artık üretildiği anda işlenerek kullanılmaktadır. Hızla oluşan yeni verileri beklemeden VM teknikleri çerçevesinde işleyerek analiz etmek, günümüz dünyası için organizasyonlara katma değer yaratmaktadır (Doğan ve Arslantekin, 2016).

VM, veri tabanlarındaki bilgiyi keşfetme-VBK (knowledge discovery in database- KDD) sürecinde, veriden yararlı unsurlar ve modeller üreten, çok önemli bir adımdır. VBK terimi ile VM terimi farklı içeriğe sahiptir. VBK, veriden yararlı bilginin keşfedilmesi aşamasının bütün sürecine işaret eder. VM, yararlı bilgiyi çekip almak için algoritmalar üzerine odaklanma yoluyla veri tabanlarındaki veri zenginliğinden yeni unsurlar keşfetmeye vurgu yapar (Fayyad, 1996).

MM ise dijital ortamda veya basılı formattaki belgelerin veya her türlü metinsel ifadenin çeşitli yöntemlerle işlenerek organizasyonlara katma değer yaratacak değerli bilgilerin çıkarılma süreçleri olarak ifade edilebilir. Bu yönüyle MM'ni; doğal dil işleme, görüntü işleme, web madenciliği veya büyük veri analizi gibi VM'nin farklı bir türü olarak değerlendirmek mümkündür.

Metin veri madenciliği veya metinde bilgi keşfi olarak da bilinen MM, genel olarak yapılandırılmamış metinden ilginç ve önemsiz olmayan bilgilerin çıkarılması sürecini ifade eder. MM, istatistik ve bilgisayar bilimleri ile bilişsel dilbilimden yararlanan disiplinlerarası bir alandır. Kütüphaneciler neredeyse dolaylı olarak MM yaparlar; onlara göre bilgi, kalıplara, gruplara, kümelere ve hiyerarşilere ayrılır. İyi bir MM uygulaması, üzerinde çalışılan bilgi tabanının kalitesine bağlı biçimde gerçekleşir. MM, artan biçimde ilgi çeken ve BY'nde aktif olarak uygulanan bir süreçtir. Veri tabanlarında yararlı gerçekleri veya bilgi parçalarını bulmak,

yapılandırılmamış metin verilerindeki gizli unsurları ortaya çıkarmaya çalışan bir analiz süreci olan MM'nin özünü oluşturmaktadır. MM, günümüzde, insan kaynakları yönetiminden pazar istihbaratına, araştırma ve geliştirmeye kadar uzanan bilgi keşfi ve iş zekası uygulamalarında kullanılmaktadır. MM teknikleri ayrıca, daha etkileşimli ve bağlamsal olarak bilinçli bir arama deneyimi yaratan özellikleri ile geleneksel bilgi erişim sistemlerinin daha geniş kitlelere ulaşması için de kullanılmaktadır (Natarajan, 2005).

Litaretüre bakıldığında genel olarak MM, metinsel verilerden bilginin çıkarılması süreci olarak tanımlanmakta ve ayrıca metin sınıflandırma, kümeleme ve ilişkilendirme kavramları, MM'nin tipik analiz süreçlerini oluşturmaktadır (Jo, 2019, s. 3). MM'nin temelini; sosyal ağlarda, internette veya veri tabanlarında yer alan, büyük miktarda ve metinlerden oluşan yapılandırılmamış veri oluşturmaktadır (Aggarwal ve Zhai, 2012, s. 1). MM'de amaç, metinlerden oluşan yapılandırılmamış veriyi yapılandırarak işlenebilecek hale getirmektedir. Metinleri yapılandırılmış veriye dönüştürme işlemiyle ifade edilmek istenen, metinlerin niceliksel olarak ölçülebilen verilere dönüştürülmesidir. Bu dönüştürme işlemini gerçekleştirebilmek için MM süreçlerinde kullanılan birtakım teknikler bulunmaktadır.

MM süreçlerindeki tekniklerden en çok kullanılanları; parçalara ayırma (tokenization), kök bulma (stemming), gereksiz kelimeleri çıkarma (stop-word removal), terim ağırlıklandırma (term weighting) ve kelime dizileri (ngram) uygulamalarıdır. Parçalara ayırma (tokenization) işlemi; metinleri, boşluk veya noktalama işaretlerine göre parçalara ayırma işlemi olarak ifade edilebilir. Örnek olarak "*metinleri, boşluk veya noktalama işaretlerine göre ayırma işlemi*" cümlesinin parçalara ayırılmış hali şu şekilde bir liste olacaktır: {metinleri, boşluk veya, noktalama, işaretlerine, göre, ayırma, işlemi}. Kök bulma (stemming) uygulaması, kelimelerin köklerine indirgenmesidir. Bir önceki örnekteki "metinleri" kelimesi ele alınacak olunursa, bu kelimenin, kelime köküne indirgemesi sonucu "metin" kelimesi elde edilecektir. Gereksiz kelimeleri çıkarma (stop-word removal) uygulaması, metin içerisindeki tek başına anlam ifade etmeyen kelime veya ifadelerin çıkarılmasıdır. Yukarıda verilen örnek cümledeki "veya" ifadesinin veri kümesinden çıkarılması bir gereksiz kelimeleri çıkarma işlemi olarak değerlendirilebilir. Terim ağırlıklandırma (term weighting) uygulaması, her bir kelimenin, tek bir metin veya veri kümesindeki tüm metinler üzerindeki ağırlığını hesaplama sürecini ifade etmektedir (Jo, 2019, s. 21-26). Kelime dizileri (ngram) yönteminde ise kelime ve kelime gruplarının ilgili metinlerde kaç defa geçtiği hesaplanmaktadır. Her metin için tek bir kelimenin frekans sayısını hesaplamaya 1-gram (unigram), bitişik iki kelimeden oluşan kelime grubunun frekansını hesaplamaya 2-gram (bigram) ve bitişik üç kelimeden oluşan kelime grubunun frekansını hesaplamaya ise 3-gram (trigram) denmektedir (Schonlau, Guenther ve Sucholutsky, 2017).

MM süreçlerinde araştırmada ele alınan probleme göre, yukarıda belirtilen işlemlerin hepsinin veya bazılarının uygulanması sonucunda kelime listeleri hazır hale getirilir. Hazırlanan listedeki kelimelerin veya kelime gruplarının ilgili dokümanda veya tüm veri kümesindeki frekansları ve/veya ağırlıkları hesaplanıp değişkenler yaratılır. Özetlemek gerekirse, her bir kelime ve/veya kelime grubu, bahsedilen teknikler kullanılarak temsili bir sayısal forma dönüştürülür, böylece analiz edilebilir yapılandırılmış (nicel) veri elde edilir.

Böylelikle elde edilen veri; sınıflama, kümeleme veya regresyon gibi çeşitli yöntemlerle işlenip analiz edilebilir.

Literatürde çok sayıda MM hakkında akademik araştırma yapıldığı görülmektedir, lakin bu araştırmada konu edinilen problemin ele alındığı başka bir araştırmaya rastlanmamıştır. Bu sebeple son yıllarda yayınlanmış bu çalışmaya yakın sayılabilecek çalışmalar incelenmiştir. Tseng (2020) MM tekniklerini kullanarak yaptığı çalışmayla internette yayınlanan haber içeriklerini analiz ederek haberlerdeki COVID-19 ile ilgili anahtar kelimeleri çıkartarak bir WordNet yapısı oluşturmuş ve bu yapının, COVID-19 salgınının sosyal ve tarihi arka planını oluşturmakta kullanılabileceğini göstermiştir. Böylelikle araştırma sonuçlarının gelecekte ortaya çıkabilecek yeni salgınlara karşı, insanlığın hızla savunmaya geçmesine yönelik içsel farkındalığını yükseltmeye yardımcı olabileceği belirtilmiştir. Choi, Shin ve Kang (2021) dijital olarak yayınlanan haber içerikleri üzerine bir araştırma yapmış ve ilgi gören içerikleri saptamada dijital platformların ve kitlelerin tercihlerinin ne denli önemli olduğunu araştırmıştır. Araştırma sonucunda haber kalitesi açısından derinlik, çeşitlilik ve inanılırlık gibi özelliklerin; okunabilirlik, nesnellik ve sansasyonellikten daha önemli parametreler olduğu ortaya çıkarılmıştır. Lyu ve Choi (2020), yaptıkları araştırmada MM analizleri kullanarak organik ürün satışı yapan çevrimiçi bir platform için, ürün satışında önemli olan faktörleri ortaya çıkarmışlardır. Araştırma sonucu ortaya çıkarılan bilgiler kullanılarak satış hacmi ve tüketici memnuniyetinin arttırılmasına yönelik yeni pazarlama stratejileri önerilmiştir. Englmeier (2021) internet üzerinde yayılan sahte haber içeriklerini tespit etmek için metin madenciliği yöntemleri kullanarak bir kontrol sistemi tasarlamıştır. Araştırma sonucu ortaya çıkarılan sistemin, internet üzerindeki yalan haber ve yanlış bilgilerin tespit edilmesini kolaylaştırması amaçlanmıştır. Tan ve diğerleri (2018) yaptıkları araştırmada, web metinlerini analiz ederek başlık bazlı bir bilgi çıkarım modeli önermişlerdir. Önerilen model, verilerdeki gürültüleri filtreleyerek web içeriklerini daha doğru konumlandırmaya/sınıflandırmaya yardımcı olması için tasarlanmıştır. Araştırma sonucu elde edilen bilgilere göre, maliyet ve zaman açısından önerilen yeni modelin, diğer benzer modellere göre daha iyi sonuçlar verdiği belirtilmiştir. Çakmak ve Eroğlu (2020) halk kütüphaneleri üzerine gerçekleştirdikleri araştırmada, belirlenen kütüphanelerin Facebook gönderilerini MM teknikleri kullanarak analiz etmişlerdir. Kütüphanelerin Facebook gönderilerinde; kullanıcı etkileşimleri, hangi konularda paylaşım yapıldığı ve gönderilerin zaman dilimleri araştırılmıştır. Araştırma sonucunda elde edilen bulgulara göre halk kütüphanelerine Facebook paylaşımları için öneriler verilmiştir.

Görülebileceği gibi MM, her türden metinsel içeriği analiz etmede sıklıkla kullanılan bir analiz yöntemidir. Bu çalışma kapsamında dijital ortamda içerik yayınlayan bir web sitesinin içerikleri MM teknikleri ile analiz edilmiştir. Uygulama, çalışmanın yöntem kısmında açıklanmış ve sonuç kısmında elde edilen bilgiler detaylandırılmış ve yorumlanmıştır.

Yöntem

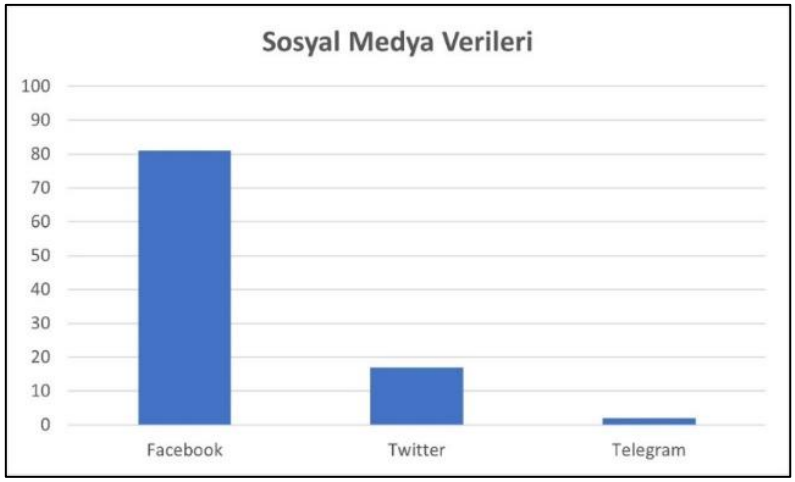
Çalışmada blockchain teknolojileri hakkında günlük olarak haber, makale, analiz ve rapor gibi içerik yayınlayan dijital bir içerik platformunun (web sitesinin) analizi yapılmıştır. İçerik analizi için bu platformun seçilme nedeni, platform verilerinin bu araştırmaya uygun olmasından kaynaklanmaktadır. Dijital içerik alanında yüksek rekabet olduğu için platform

isminin açıklanmasını istememiştir. Platformun içerikleri Türkçe olarak yayınlanmaktadır ve platformda yayınlanan içerikler aylık beş milyondan fazla okunmaktadır.

Bu araştırma, seçilen web sitesinin sosyal medyada paylaştığı içeriklerin başlık bazında incelenmesine dayanmaktadır. Web sitesi yayınladığı tüm içerikleri, sosyal medya hesaplarından da paylaşmaktadır. Şirketin kendi Twitter, Telegram ve Facebook hesaplarından paylaşılan haber, rapor, makale, analiz ve değerlendirme yazıları gibi çeşitli içerikler, siteye okuyucu çekme amacı taşımaktadır. Bu nedenle sosyal medya hesaplarından ilgili başlıklara tıklayarak siteye gelen okuyucu kitlesi, organizasyon için önemli görülmektedir. Platformun sosyal medyadan gelen istatistikleri incelendiğinde, Şekil 1'de görülebileceği gibi, sosyal medya mecralarından gelen okuyucuların %80'i Facebook kanalıyla platforma erişmektedir.

Şekil 1

Platformun sosyal medya verileri



Bu çerçevede çalışmada analiz için platformun en büyük sosyal medya kullanıcı kaynağı olan Facebook seçilmiştir. Araştırmada Facebook'ta gönderi olarak paylaşılan içeriklerin başlıkları analiz edilerek, başlık bazında okunma oranını etkileyen kelime ve kelime gruplarının belirlenmesi amaçlanmıştır.

Facebook gibi sosyal medya platformlarında oluşturulan gönderilerde, kullanıcılara içeriklerin sadece başlıkları gösterilmekte ve içeriğin kendisi tıklama yapıldıktan sonra ilgili sayfaya yönlendirme yapılarak okunabilmektedir. Bundan dolayı bir kullanıcının o içeriğe tıklayıp tıklamaması, başlığın okurun ilgisini çekip çekmemesine ve sonuç olarak başlıkta kullanılan kelime ve kelime gruplarına bağlıdır. Bu nedenlerle çalışmada, istenen sonuca ulaşmak için yayınlanan içeriklerde başlık bazında okunma oranını etkileyen faktörler (kelime ve kelime grupları) üzerine analizler yapılmıştır. Bu bağlamda, bu araştırmada iki farklı analiz yapılmış ve sonuçlar karşılaştırılmıştır. İlk analizde araştırmanın ele aldığı probleme özel olarak farklı bir yaklaşım uygulanmış, ikinci analizde ise standart MM teknikleri kullanılmıştır.

Verilerin Hazırlanması

Araştırmada, ham veri üzerinde düzenlemelerin yapılabilmesi için Excel programı; MM süreçleri, değişken yaratma ve veri ön işleme işlemleri için Python programlama dili; regresyon

analizi ve istatiksel testler için Minitab istatistik yazılımı kullanılmıştır. Verilerin hazırlanmasında Numpy, Pandas, Seaborn, Matplotlib, Statsmodels, SnowballStemmer ve Sklearn veri kütüphaneleri kullanılarak JubiterLab platformu üzerinde Pyhton programlama diliyle kodlama yapılmıştır.

Metin Ön İşleme

Analizin uygulanacağı veri kümesini oluşturmak için üç aylık bir dönem belirlenmiş ve bir sınırlama yapılmıştır. Bunun için 01 Mart 2022 ile 31 Mayıs 2022 tarih aralığı seçilmiş ve bu tarihler arasında Facebook vasıtasıyla platforma gelen oturum istatistikleri, ilgili şirketten Excel dosyası olarak sağlanmıştır. Elde edilen veri kümesi 2206 gözlemden oluşmaktadır. Her bir gözlem, benzersiz bir içerik başlığı (metin) ve oturum istatistiğinden oluşmaktadır. Oturum istatistiği, kullanıcıların Facebook gönderileri vasıtasıyla içeriklere tıklayarak web sitesine kaç kez gelindiğini göstermektedir.

2206 satırdan oluşan Excel dosyası üzerinde =RASGDİZİ() fonksiyonu kullanılarak oransız eleman örnekleme yöntemiyle rastgele seçilen 500 içeriğe ait veriler, ayrı bir dosya olarak kaydedilmiştir. Kaydedilen Excel dosyası, araştırmanın veri kümesini oluşturmaktadır.

Oransız eleman örnekleme yöntemi, çalışma evrenindeki gözlemlerin eşit oranda seçilme ihtimaline sahip olduğu yansız bir örnekleme türüdür. Çalışma evrenini tamamıyla veya gereğinden büyük bir örneklem ile analiz etmek, değişkenlerin kontrollerini güçleştirebilmektedir. Araştırmalarda temel hedef, çok veri ile analiz yapmak değil, geçerli ve tutarlı verileri işleyebilmektir (Karasar, 2022, s. 148, 151). Verilerin rastgele dağılıma sahip olduğunu kabul eden parametrik modellerde (bu çalışmada uygulanan doğrusal regresyon) büyük verilerin modellenmesinde gerekli varsayımların sağlanamadığı bilinmektedir (Batmaz ve diğerleri, 2017; Batmaz, Karagöz ve Serdar, 2017). Bu nedenle literatürde büyük veri karmaşasının üstesinden gelmek için genel olarak iki yaklaşım kullanılmaktadır; örnekleme ve dağıtık sistemlerin kullanımı (Bifet, 2013). Örnekleme yöntemiyle belleğin küçüklüğünü ve zaman kısıtlılığını göz önünde bulundurarak analiz için örnek veri kümesi elde edilebilir (Batmaz ve diğerleri, 2017; Altınok, Karagöz ve Batmaz, 2021). Öte yandan parametrik olmayan yöntemlerle de veri modellenebilir. Ancak bu çalışmada parametrik modeller kullanılmak istenildiğinden örnekleme yaklaşımı uygulanmıştır. Sonraki çalışmalarda bu çalışmanın ele aldığı problem için tüm veri kümesi kullanılarak parametrik olmayan modeller ele alınabilir. Bu nedenlerle bu araştırma, metin madenciliği teknikleriyle yaratılan, büyük miktarda değişken grupları içerdiğinden, oransız eleman örnekleme yöntemi uygulanmıştır. Örnekleme yöntemi uygulanarak değişken sayısı kontrol altında tutulmak istenmiş ve model varsayımlarının sağlanabildiği bir veri kümesi oluşturmak hedeflenmiştir.

Elde edilen Excel formatındaki dosya, JupyterLab uygulamasına aktarılarak metin ön işleme için kodlama yapılmıştır. Bu aşamada, içerik başlıklarında büyük-küçük harf dönüşümü yapılarak tüm karakterler küçük harfe dönüştürülmüş, noktalama işaretleri ve özel karakterler kaldırılmış ve son olarak rakamlar temizlenmiştir.

Metin ön işleme aşamasından sonra elde edilen veri kümesi Tablo 1'de örnek olarak verilmiştir. Başlık ve oturum sayısından oluşan ham veri kümesi, iki sütun ve 500 satırdan oluşmaktadır.

Tablo 1

Veri kümesi

Başlık	Oturum
bitcoin neden düşüyor btc düşüşünün arkasındaki faktör	191
nftimi kiraya vermek istiyorum bebek fluffyler geliyor	88
altcoin kurucusunun hesabı açıldı kripto para yatırımcılarına uyarılar	258

Not. Oturum sütunundaki değerler Facebook kanalıyla siteye gelen oturumların sayısını göstermektedir.

Değişken Yaratma

Çoğu metin madenciliği çalışmasında, kelimeleri köke indirgeme (stemming) ve tek başına anlam ifade etmeyen "ve", "veya", "bu" gibi edat, zamir veya bağlaç ifadelerini kaldırma (stop-word removal) işlemi yapılmaktadır. Çalışmanın birinci analizinde bu işlemler yapılmadan kelimeler farklı bir yaklaşımla ele alınmıştır. Bu farklı yaklaşımın uygulanmasındaki amaç, her bir kelimenin, kendisini hiçbir değişiklik yapmadan ele alarak okunma oranı üzerindeki etkisini ortaya çıkarmaktır. Çünkü kullanıcılar, internet üzerinde yayınlanan haber, analiz, rapor, duyuru vb. gibi içeriklerin ilk önce başlıklarını okuyarak bir etkileşime girerler. Ayrıca kullanıcılar bir içeriğe tıklayıp tıklamama kararını başlığı okuyarak hızlıca verebilmektedir. Bu nedenle yazarlar tarafından başlıklarda kullanılan her bir kelime ve kelime grubunun nasıl kullanıldığı çok önemlidir. Bu bağlamda ele alındığında, ilk etapta tek başına bir anlam ifade etmiyor gibi görünen bir kelime, başka kelimelerle bir araya gelerek önceden tahmin etmesi mümkün olmayan anlamlı sonuçlar verebilir. Bu nedenlerle ilk analizde kelimeler olduğu gibi ele alınmıştır. JubiterLab üzerinde kodlama yapılarak başlık sütunundaki her bir metin, kelime ve kelime gruplarına ayrılmış ve frekans sayılarına göre listelenmiştir. Kelimelerin 1-gram (unigram), 2-gram (bigram) ve 3-gram (trigram) olmak üzere terim frekansları çıkarılmıştır. Bunun için kelime dizileri (ngram) yöntemi kullanılmıştır. Örnek olarak verilen "bitcoin neden düşüyor btc düşüşünün arkasındaki faktör" şeklindeki başlıkta kelime ve kelime grupları, aşağıda gösterildiği gibi çözümlenmiştir. Örnek olarak verilen bu başlıktan kelime dizileri (ngram) tekniği kullanılarak 18 adet özellik (değişken) elde edilebilir.

1-gram: "bitcoin", "neden", düşüyor", "btc", "düşüşünün", "arkasındaki", faktör"

2-gram: "bitcoin neden", "neden düşüyor", "düşüyor btc", "btc düşüşünün", "düşüşünün arkasındaki", "arkasındaki faktör"

3-gram: "bitcoin neden düşüyor", "neden düşüyor btc", "düşüyor btc düşüşünün", "btc düşüşünün arkasındaki", "düşüşünün arkasındaki faktör"

Yapılan değişken yaratma çalışmaları sonucunda, ilk analiz için toplamda 7883 benzersiz kelime ve kelime grubu tespit edilmiştir. Toplam benzersiz kelime ve kelime grubu sayısı 7883 olmasına rağmen, bunların 6890 tanesi tüm veri kümesi içerisinde tek sefer geçmektedir, yani frekans sayısı birdir. Bu çalışmanın da konu edindiği parametrik modellerde varsayımların sağlanabilmesi ve istatistiksel olarak geçerli bir analiz gerçekleştirebilmek için değişken sayısı kontrol altında tutulmak istenmiştir. Çünkü ortaya çıkan 7883 benzersiz kelime ve kelime grubu (değişken) ile istatistiksel olarak geçerli bir parametrik model kurmak; işlemci gücü ve bellek kısıtları bakımından mümkün görünmemektedir. Parametrik olmayan modeller ile tüm kelimeler

dahil edilerek yapılacak bir araştırma sonraki çalışmalarda ele alınabilir. Bu çalışmada iki farklı analiz için aynı örneklem veri kümesi ve aynı frekans sayıları kullanıldığından, iki analiz sonucunu tutarlı bir şekilde karşılaştırabilmek mümkün olmaktadır. Bu nedenlerle frekans sayısı iki ve üzeri olan 993 kelime ve kelime grubu analize dahil edilmiştir.

Tablo 2'de birinci analiz için frekansa sayısına göre sıralanmış ilk 10 kelime ve kelime grubu gösterilmektedir.

Tablo 2

Birinci analiz için frekans sayısına göre ilk 10 kelime ve kelime grubu

Sıra	Kelime ve kelime grubu	Frekans	Sıra	Kelime ve kelime grubu	Frekans
1	bu	154	6	son	77
2	ve	117	7	para	54
3	kripto	90	8	btc	54
4	bitcoin	80	9	kripto para	51
5	altcoin	77	10	durum	45

Not. Kelimelerin başlıklarda geçme sıklığına göre sıralama yapılmıştır.

İkinci analizde kullanılmak üzere standart MM tekniklerinden olan kelimeleri köke indirgeme (stemming) ve tek başına anlam ifade etmeyen kelimeleri kaldırma (stop-word removal) işlemi yapılarak veri kümesi hazırlanmıştır. Bu işlemler için Türkçe dilini destekleyen SnowballStemmer1 uygulaması kullanılmıştır. SnowballStemmer algoritması vasıtasıyla ilgili kodlamalar yapılarak tüm kelimeler köklerine indirgenmiş ve tek başına anlam ifade etmeyen kelimeler (stop-words) kaldırılmıştır.

Sonrasında ilk analiz için oluşturulan veri kümesi için aynı işlemler tekrarlanmış ve kelimelerin 1-gram (unigram), 2-gram (bigram) ve 3-gram (trigram) olmak üzere terim frekansları çıkarılmıştır. İlk analizin veri kümesiyle ikinci analizin veri kümesi arasında karşılaştırma yapabilmek için tutarlılık sağlamak amacıyla aynı şekilde uygulama yapılmış ve frekans sayısı iki ve üzerinde olan 844 kelime araştırmaya dahil edilmiştir.

Tablo 3'te ikinci analiz için frekansa sayısına göre sıralanmış ilk 10 kelime ve kelime grubu görülebilir.

Tablo 3

İkinci analiz için frekans sayısına göre ilk 10 kelime ve kelime grubu

Sıra	Kelime ve kelime grubu	Frekans	Sıra	Kelime ve kelime grubu	Frekans
1	altcoin	147	6	para	54
2	kripto	99	7	kripto para	51
3	bitcoin	95	8	fiyat	44
4	son	80	9	son durum	42
5	btc	54	10	bitcoin btc	41

Not. Kelimelerin başlıklarda geçme sıklığına göre sıralama yapılmıştır.

[1] SnowballStemmer Türkçe dilini destekleyen, doğal dil işlemede kullanılan bir algoritmadır.

Standartlaşmış MM yaklaşımıyla ve çalışmanın ele aldığı probleme özel olarak uygulanan farklı bir yaklaşımla iki ayrı veri kümesi hazırlanmıştır. Tablo 2 ve Tablo 3'de iki ayrı veri kümesi üzerindeki farklılıklar görülebilir.

Tablo 2'de "bu", "ve" gibi kelimeler varken, Tablo 3'de bu kelimeler bulunmamaktadır. Ayrıca ikinci analiz için hazırlanan veri kümesinde kelimeler köklerine indirgendiğinden dolayı bazı kelimelerin frekans sayıları, diğer veri kümesine göre daha yüksektir. Bir örnekle açıklamak gerekirse, "kripto" kelimesi Tablo 2'de 90 frekansa sahipken, Tablo 3'te 99 frekansa sahip olarak hesaplanmıştır. Bunun nedeni kelimeye eklenen eklerin atılmasından dolayıdır. Metinler içerisinde yer alan "kriptolar" veya "kriptoya" gibi ek almış kelimeler, ekleri atıldığında "kripto" kelime kökü altında toplanmıştır, bu nedenle "kripto" kelimesinin Tablo 3'te frekans sayısı daha yüksektir.

Her iki yaklaşımla listelenen kelime ve kelime grupları, JubiterLab üzerinde ilgili kodlamalar yapılarak vektör haline getirilmiş ve her biri, bir değişken olarak kaydedilmiştir. Bu işlemler sonrası iki analiz için iki ayrı veri kümesi uygun hale getirilmiştir.

Her iki yaklaşımla elde edilen veri kümelerine bir sonraki bölümde detaylarıyla açıklanan çoklu doğrusal regresyon yöntemi uygulanmıştır.

Analiz

Farklı bir yaklaşımla veri kümesinin hazırlandığı birinci analiz için, içeriklerin okunma sayılarını ifade eden oturum değişkeni (y) bağımlı değişken ve oluşturulan 993 değişken ise bağımsız değişkenler olarak atanmış ve çoklu doğrusal regresyon yöntemi ile Minitab programında analiz edilmiştir. 993 bağımsız değişkenden hangilerinin modele anlamlı bir katkı sunduğunu ortaya çıkarmak için iki aşamalı bir strateji uygulanmıştır.

İki aşamalı strateji değişken sayısının çok fazla olduğu durumlarda kullanılmaktadır. İki aşamalı stratejide öncelikle adımsal regresyon yöntemleri uygulanarak göz ardı edilebilir etkileri olan bağımsız değişkenler çıkarılır ve değişken sayısı büyük oranda azaltılır, daha sonra azaltılmış değişkenler kümesiyle tekrar analiz yapılarak nihai modele ulaşılabilir (Montgomery, Peck ve Vining, 2013, s. 353).

Bu çalışmada iki aşamalı stratejiyi uygulamak için ilk aşamada adımsal (stepwise regresyon), ikinci aşamada ise geriye doğru eleme (backward regresyon) yöntemleri uygulanmıştır.

Adımsal (stepwise) regresyon yönteminde, her adımda başta belirlenen güven aralığına göre p-değeri en küçük değişken modele dahil edilir. Yeni değişkenin modele dahil edilmesi ile birlikte model içerisinde yer alan değişkenlerden p-değeri (belirlenen güven aralığına göre) yeterince büyük olanlar modelden çıkarılır. Uygulama regresyon modeline dahil edilecek veya çıkarılacak değişken kalmayıncaya kadar sürer (Pektaş, 2013, s. 122).

Geriye doğru eleme (backward regresyon) yöntemi ise modeldeki tüm aday değişkenlerle başlar. Her adımda, belirlenen anlamlılık düzeyine göre en az anlamlı olan değişken çıkarılır ve bu işlem anlamsız değişken kalmayana kadar devam eder (Trzepiecinski, Szpunar ve Kascak, 2021).

$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_p x_{pj} + \varepsilon_{ji}$ denkleminde çoklu doğrusal regresyon eşitliği gösterilmektedir (Albayrak, 2006, s. 225).

İki aşamalı stratejinin ilk aşamasında, %95 güven aralığı (p-value < 0,05) belirlenmiş ve adımsal regresyon yönteminde denkleme giriş için α (alpha) değeri 0,05, çıkış için 0,10 değeri kullanılmıştır. Bu aşamadaki amaç sadece değişken sayısını azaltmak olduğu için alpha değerleri geniş tutulmuştur. Adımsal regresyon yöntemi ile başlangıçta 993 adet bağımsız değişken varken, analiz sonucu modele anlamlı katkı sunabilecek 131 bağımsız değişken saptanmıştır. Fakat bunları doğrudan kullanmak tutarlı sonuçlar vermemektedir, çünkü doğrusal regresyon analizinde bazı varsayımlar bulunmaktadır.

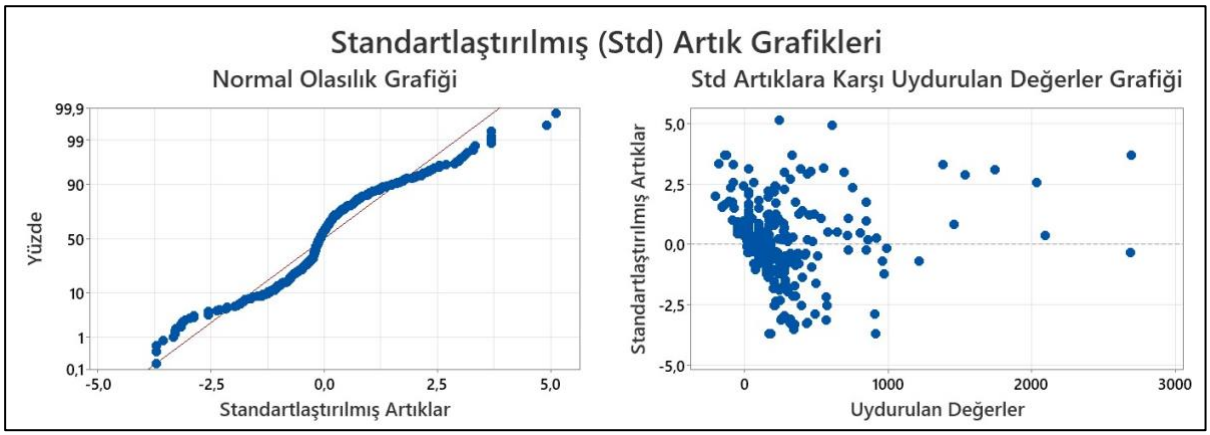
Doğrusal regresyon analizinde varsayımların sağlanması, doğru ve geçerli bir analiz yapılması için oldukça önemlidir. Varsayımları sağlanamayan modeller, yetersiz olup kararsız bir model ortaya çıkarabilecek ciddi sonuçlara sahiptir (Montgomery, Peck ve Vining, 2013, s. 129).

Doğrusal regresyon beş temel varsayıma dayanmaktadır. Varsayımlar şu şekilde özetlenebilir: 1) Hatalar normal dağılır; 2) Hataların varyansları sabittir; 3) Hatalar bağımsızdır ve aralarında otokorelasyon yoktur; 4) Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki doğrusaldır; 5) Bağımsız değişkenler arasında çoklu doğrusal bağlantı problemi yoktur (Albayrak, 2006, s. 49).

Şekil 2'de adımsal regresyon yöntemiyle ortaya çıkan artık dağılımları gösterilmektedir. Normal olasılık grafiğinde bazı standartlaştırılmış artıkların referans doğrusudan kaydığı gözlemlenmektedir, ayrıca std artıklara karşı uydurulan değerler grafiği incelendiğinde, sola çarpık bir dağılım gözlemlenmektedir. Bu durum veri kümesinde bazı problemler olduğuna ve doğrusal regresyon varsayımlarının sağlanamadığına işaret etmektedir.

Şekil 2

Standartlaştırılmış Artık Grafikleri



İlk aşama analizinde, geçerli ve tutarlı olabilecek bir değişken kümesi oluşturabilmek için varsayımların minimal düzeyde de olsa sağlanması amaçlanmıştır, çünkü nihai model ikinci aşamada kurulmuştur. Bu aşamada büyük değişken kümesinden hiç etkisi olmayan değişkenleri çıkartarak daha küçük bir değişken kümesi elde etmek hedeflenmektedir. Bu sebeplerle ilk aşamada daha tutarlı sonuçlar alınabilmesi için bağımlı değişken olan oturum değişkenine Box-Cox dönüşümü yapılmıştır.

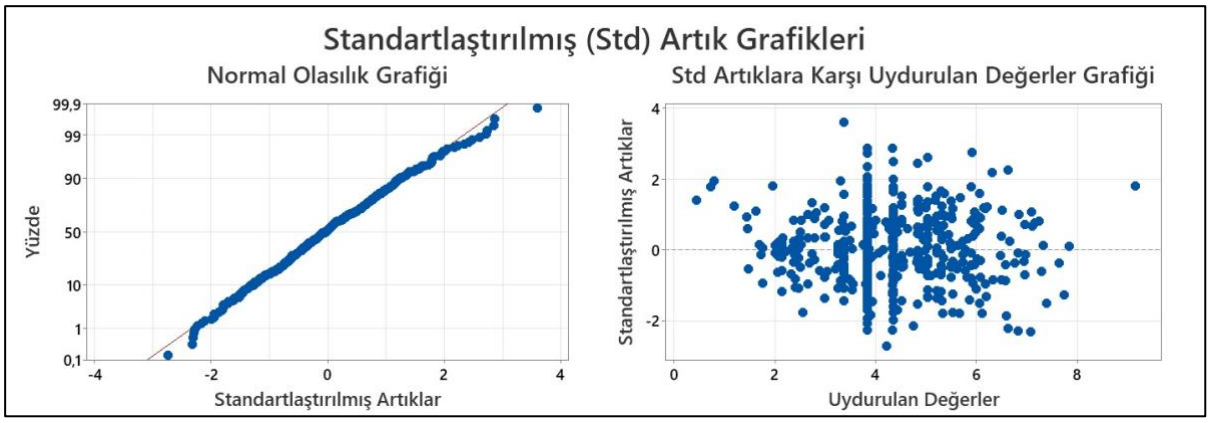
Box-Cox dönüşümü, verideki eş varyanslık ve normalliği geliştirmek için uygulanan bir parametrik dönüşüm tekniğidir. Bu dönüşüm sayesinde sağa veya sola çarpık dağılımlar optimize edilir (Xiong ve diğerleri, 2021).

Box-Cox dönüşümü yapıldıktan sonra %95 güven aralığı (*p*-value < 0,05) belirlenmiş ve adımsal regresyon yönteminde denkleme giriş için α (alpha) değeri 0,05, çıkış için 0,10 değeri kullanılmıştır. Başlangıçta 993 adet bağımsız değişken varken, analiz sonucu modele anlamlı katkı sunabilecek 72 bağımsız değişken saptanmıştır.

Şekil 3'te analiz sonucu ortaya çıkan artık dağılımları gösterilmektedir. Normal olasılık grafiğinde standartlaştırılmış artıkların referans doğrusunu takip ettiği gözlemlenmektedir. Ayrıca std artıklara karşı uydurulan değerler grafiği incelendiğinde, Şekil 2'dekine göre sola çarpık olan dağılımın Şekil 3'te normalleştiği görülmektedir. Bu nedenlerden dolayı Box-Cox dönüşümünün veri kümesine uygun olduğu söylenebilir. Bu aşamada belirlenen 72 değişken, ikinci aşama analiz için değişken kümesini oluşturmuştur.

Şekil 3

Standartlaştırılmış Artık Grafikleri

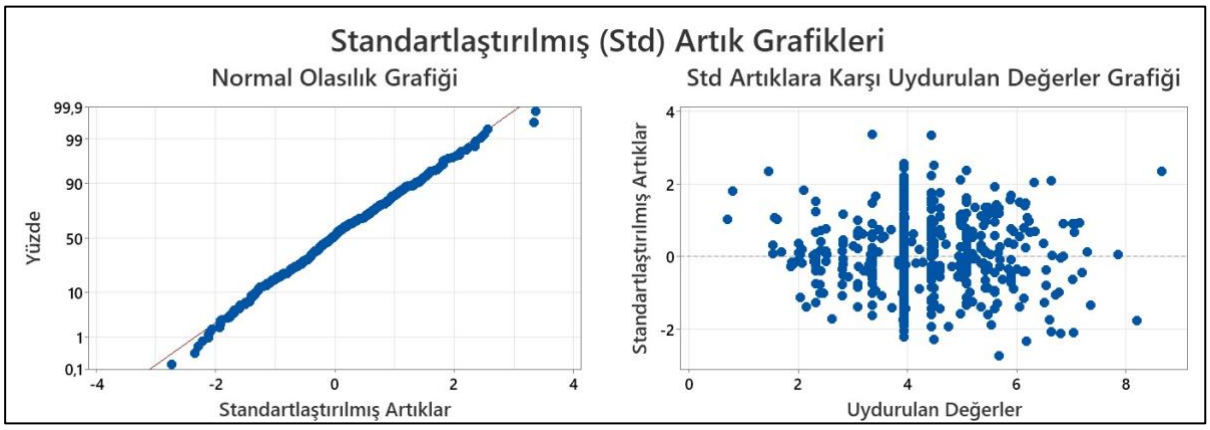


İkinci aşamada nihai modelin oluşturulması için ilk aşamada belirlenen 72 bağımsız değişken kullanılmıştır. Box-Cox dönüşümü yapılmış bir bağımlı değişken, 72 bağımsız değişken ile geriye doğru eleme (backward regresyon) yöntemi uygulanmıştır. Nihai modeli oluşturacak olan bu analiz için alpha değeri, rastsallığı azaltmak ve daha güvenilir bir model yaratabilmek için 0,025'e düşürülmüştür. %97,5 güven aralığıyla (*p*-value < 0,025) geriye doğru eleme yönteminde denklemden çıkış için α (alpha) değeri olarak 0,025 kullanılmıştır.

Başlangıçta 72 adet bağımsız değişken varken, analiz sonucu modele anlamlı katkı sunabilecek 53 bağımsız değişken saptanmıştır. Şekil 4'teki normal olasılık grafiği incelendiğinde, standartlaştırılmış artıkların referans doğrusunu takip ettiği görülmektedir. Ek olarak sdt artıklara karşı uydurulan değerler grafiği ele alındığında, çarpık bir dağılım gözlemlenmemiştir.

Şekil 4

Standartlaştırılmış Artık Grafikleri



Nihai modeli oluşturacak olan bu analizin geçerli ve tutarlı olabilmesi için, öncesinde bahsedilen beş temel doğrusal regresyon varsayımının karşılanması gerekmektedir. Bu nedenle tüm doğrusal regresyon varsayımları sınanmıştır. İlk olarak normallik varsayımı test edilmiştir.

Ryan-Joiner (RJ) normallik testi, uygulama verileri ile verilerin normal skorları arasındaki korelasyonu hesaplayarak verilerin normal dağılıp dağılmadığını ölçmektedir. Korelasyon katsayısı bire yakınsa verilerin normal dağıldığı söylenebilir (Nosakhare ve Bright, 2017).

"Hatalar normal dağılır" varsayımının sağlanıp sağlanamadığını kontrol etmek için RJ testi yapılmıştır. Normallik testi için boş ve alternatif hipotez[2] yaratılmıştır.

- H0: Hatalar normal dağılım göstermektedir.
- H1: Hatalar normal dağılım göstermemektedir.

RJ test sonucuna göre *p*-değeri 0,10 olarak bulunmuştur. Bu değer belirlenen α =0.025 değerinden büyük olduğu için, hataların normal dağılım gösterdiğini belirten H0 hipotezi reddedilmez. Sonuç olarak, hataların normal dağıldığını ifade eden varsayım sağlanmaktadır.

"Hataların varyansları sabittir" varsayımı için Şekil 4'te yer alan std artıklara karşı uydurulan değerler grafiği incelenmiştir ve "hataların varyansları sabittir" şeklindeki ifadeyi yanlışlayacak bir kanıt bulunamamıştır.

Doğrusal regresyon varsayımlarından bir diğeri olan "hatalar bağımsızdır ve aralarında otokorelasyon yoktur" varsayımını test etmek için Durbin-Watson (DW) istatistiği yöntemi kullanılmıştır.

DW istatistiği, sıfır ile dört arasında bir değer almaktadır. İkiye yakın bir değer otokorelasyon olmadığını gösterirken, sıfıra doğru bir değer pozitif otokorelasyonu; dörde doğru hesaplanan bir değer ise negatif otokorelasyonu göstermektedir (Kanji, 2006, s. 169). DW istatistiğiyle ilgili aşağıdaki gibi bir çıkarımda bulunmak mümkündür.

- $d < dL$ ise pozitif otokorelasyon vardır,
- $dL < d < dU$ ise pozitif otokorelasyonun varlığı hakkında yeterli kanıt yoktur (test sonuçlandırılamamaktadır),

[2] Doğrusal regresyon varsayımlarının kontrol edilmesi için hipotezler oluşturulmuş ve varsayımlar test edilmiştir.

- dU < d < 4-dU ise otokorelasyon yoktur,
- 4-dU < d < 4-dL ise negatif otokorelasyonun varlığı hakkında yeterli kanıt yoktur (test sonuçlandırılamamaktadır),
- 4-dL < d ise negatif otokorelasyon vardır (Uysal ve Günay, 2001).

Kullanılan kısaltmalar açıklanacak olunursa; d analiz için hesaplanan DW istatistiğini, dL (d lower) ve dU (d upper) ise DW istatistiğinin alt ve üst kritik tablo değerlerini belirtmektedir. Bu kritik değerler modeldeki gözlem sayısı ve değişken sayısına göre değişim göstermektedir (Albayrak, 2006, s. 234).

Testi gerçekleştirmek için oluşturulan hipotezler:

- H0: Hatalar arasında otokorelasyon yoktur.
- H1: Hatalar arasında negatif otokorelasyon vardır.

Minitab'da DW istatistiği hesaplanmış ve 2,05 (d) değeri bulunmuştur. Çalışmada n=500 ve k=53 (n veri kümesindeki gözlem sayısı, k ise modeldeki bağımsız değişken sayısıdır) olduğundan DW tablosunda[3] veri kümesindeki değerlere en yakın olan dL=1,46 ve dU=1,89 değerleri değerlendirmeye alınmıştır. 1,89 < 2,05 < 2,11 değerleri ile dU < d < 4-dU koşulu sağlandığından dolayı H0 hipotezinin reddedilemeyeceği söylenebilir. Sonuç olarak hatalar arasında otokorelasyon yoktur ve varsayım sağlanmaktadır.

"Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki doğrusaldır" varsayımı için Şekil 4'te yer alan std artıklara karşı uydurulan değerler grafiği incelenmiştir. Grafik incelendiğinde modelde uyum sorunu gözlemlenmemiştir.

Doğrusal regresyonun son varsayımı olan "bağımsız değişkenler arasında çoklu doğrusal bağlantı problemi yoktur" varsayımı için VIF (Variance Inflation Factors) değerleri incelenmiştir.

Variance Inflation Factors (VIF), farklı model yapıları ile doğrusallığı ölçmek ve analiz etmek için bir değerlendirme aracı olarak kullanılmaktadır. VIF, değişken çiftleri yüksek bir korelasyona sahip olmasa bile, değişkenler arasındaki doğrusallığı değerlendirebilen iyi bir değerlendirme yöntemidir. 10'dan büyük olan VIF değerleri bağımsız değişkenler arasında yüksek korelasyona işaret etmektedir (Wang ve Zou, 2018). Modeldeki VIF değerleri incelendiğinde tümünün bire yakın olduğu gözlemlenmiştir.

Doğrusal regresyon varsayımları dışında veri kümesindeki aykırı ve etkili gözlemler incelenmiştir.

Standartlaştırılmış artıklarda -3 ve 3 arası dışında kalan değerler potansiyel olarak aykırı gözlem olarak değerlendirilebilir (Montgomery, Peck ve Vining, 2013, s. 131). Analiz sonucu elde edilen standartlaştırılmış artıklar incelendiğinde, -3 ve 3 arası dışında kalan 2 adet gözlem tespit edilmiştir ve bu gözlemler aykırı olarak değerlendirilebilir.

Modeldeki etkili gözlemleri tespit etmek için ise hii, Cook Distance ve DFFITS olmak üzere üç farklı yöntem uygulanmıştır.

[3] DW istatistiğinin alt ve üst kritik tablo değerleri için Savin ve White'ın (1977) DW tablosu kullanılmıştır.

Kaldıraç noktası (hii), regresyon doğrusu üzerinde yer alan fakat örneklem noktalarının kalan kısmından uzakta olan bir konumdadır. Bu nokta olağan dışı bir değere sahiptir ve modelin belirli özelliklerini kontrol edebilir. Bu nedenle incelenmeleri önemlidir. Kaldıraç noktasına ilişkin sınır değer 2*p/n formülü ile hesaplanmaktadır. p modeldeki sabit katsayı dahil bağımsız değişkenlerin sayısıdır, n ise veri kümesindeki gözlem sayısıdır (Montgomery, Peck ve Vining, 2013, s. 212). Kaldıraç noktasına ait sınır değer 2*54/500=0,216 olarak hesaplanmıştır. Veri kümesi için hesaplanan hii değerleri incelendiğinde 0,216 noktasından büyük olan 86 adet potansiyel etkili gözlem tespit edilmiştir.

Cook Distance ve DFFITS etkili gözlem ölçüleri ise regresyon doğrusu üzerinde yer almayan uzak gözlemlerdir. Bu gözlemler model katsayıları üzerinde önemli bir etkiye sahiptir ve modeli kendi yönüne doğru çeker. Bu nedenle etkili gözlemler dikkatle incelenmelidir. Cook Distance etkinlik ölçüsünde $D_i > 1$ koşulunu sağlayan gözlemler etkili olarak değerlendirilebilir (Montgomery, Peck ve Vining, 2013, s. 211, 216). Etkinlik ölçülerinden bir diğeri olan DFFITS'e ait sınır değer, |DFFITS| > 1 koşulu ile tespit edilmektedir (Albayrak, 2006). Analiz sonucu hesaplanan Cook (D) değerleri incelendiğinde $D_i > 1$ koşulunu sağlayan herhangi bir gözlem bulunamamıştır. DFFITS tablo değerleri incelendiğinde ise 26 adet gözlem DFFITS sınır değerinden büyüktür ve potansiyel etkili gözlem olarak değerlendirilebilir.

Analiz sonucu ortaya çıkan potansiyel etkili gözlem olarak etiketlenen gözlemler içerisinden, üç etkinlik ölçüsünün (hii, Cook ve DFFITS) herhangi ikisinde eşik değerleri geçen gözlemler etkili gözlem olarak değerlendirilir (Batmaz, 2021).

Yapılan iyileştirme çalışmaları ve testler sonucunda tüm doğrusal regresyon varsayımları sağlanmıştır. Veri kümesinde 2 adet aykırı gözlem bulunmakla birlikte; hii, Cook ve DFFITS değerlerinden herhangi ikisinde eşik değerleri geçen 26 adet etkili gözlem bulunmaktadır. Etkili ve aykırı gözlemlerin model sonuçlarına etki edebileceği düşünüldüğünden, istatistiksel olarak tutarlı bir sonuç elde edebilmek için, etkili gözlemler (26 adet) ve aykırı gözlemler (2 adet) veri seti dışına alınmıştır. Kalan 472 gözlem ile geriye doğru eleme yöntemi tekrar uygulanmıştır. %97,5 güven aralığıyla (p-value < 0,025) denklemden çıkış için α (alpha) değeri olarak 0,025 kullanılmıştır ve varsayım testleri tekrarlanmıştır. Sonuç olarak, 472 gözlem ile yapılan bu son analizde önerilen model için tüm doğrusal regresyon varsayımları sağlanmıştır. Ayrıca modelde aykırı değer bulunmamaktadır ve üç etkinlik ölçüsünün herhangi ikisinde eşik değerleri geçen gözlem saptanmamıştır. Analiz sonucunda %97,5 güven aralığı (p-value < 0,025) ile modele anlamlı katkı sunan 39 bağımsız değişken saptanmıştır. Yapılan analizde elde edilen model için tüm varsayımlar sağlandığından ve veri kümesinde aykırı ya da etkili gözlem olmadığından, oluşturulan model kullanıma uygun bulunmuştur. Bulgular kısmında birinci analizin model detaylarına yer verilmiştir.

Standart MM teknikleriyle veri kümesinin hazırlandığı ikinci analiz için, içeriklerin okunma sayılarını ifade eden oturum değişkeni (y) bağımlı değişken ve oluşturulan 844 değişken ise bağımsız değişkenler olarak atanmış ve çoklu doğrusal regresyon analizi yapılmak üzere Minitab programında analiz edilmiştir. İkinci analizde, birinci analizde yapılan işlemlerin aynısı uygulanmıştır. Tekrar olmaması açısından ikinci analiz özet olarak verilmiştir.

İkinci analizde doğrusal regresyon yönteminin ilk aşaması için %95 güven aralığı (p-value < 0,05) belirlenmiş ve ilk analizdeki gibi adımsal regresyon (stepwise regresyon) yönteminde denkleme giriş için α (alpha) değeri 0,05, çıkış için 0,10 değeri kullanılmıştır. Elde

edilen sonuçlar değerlendirilerek bağımlı değişkene Box-Cox dönüşümü uygulanmış ve model tekrar çalıştırılarak başlangıçtaki 844 değişkenden istatistiksel olarak anlamlı bulunan 51 değişken belirlenmiştir. İkinci aşamada, belirlenen 51 değişken geriye doğru eleme (backward regresyon) yöntemiyle ve %97,5 güven aralığıyla (p-value < 0,025), denklemden çıkış değeri olan 0,025 kullanılarak uygulanmıştır.

Sonuçlar incelendiğinde 2 adet aykırı ve 24 adet etkili gözlem tespit edilmiştir. Bu gözlemler veri kümesi dışında bırakılarak analiz 474 gözlemle tekrarlanmış ve model için istatistiksel olarak anlamlı olan 30 değişken saptanmıştır. Elde edilen model için varsayım testleri yapılmış ve tüm varsayımlar sağlanmıştır. Ayrıca modelde aykırı ve etkili gözlem bulunmamaktadır. Yapılan analizde elde edilen model için tüm varsayımlar sağlandığından ve veri kümesinde aykırı ya da etkili gözlem olmadığından, oluşturulan model kullanıma uygun bulunmuştur. Bulgular bölümünde standart MM teknikleriyle uygulanan ikinci analizin model detaylarına yer verilmiştir.

Bulgular

Tablo 4'te farklı bir yaklaşımla ele alınan birinci analizin modeline yer verilmiştir. Bu tabloda 39 değişkene ait katsayılar, p-değerleri ve VIF değerleri gösterilmektedir. Birinci modelin R^2 değeri %62,68, R^2_{adj} (düzeltilmiş R^2) değeri ise %59,31 olarak hesaplanmıştır. Araştırma sonuçlarında R^2_{adj} (düzeltilmiş R^2) değeri esas alınmıştır.

Tablo 4

Birinci analizin model özeti

Değişkenler	Katsayılar	p-değeri	VIF	Değişkenler	Katsayılar	p-değeri	VIF
Constant (sabit)	3,9098	0,000		inu	1,049	0,000	1,11
altcoin	0,664	0,000	1,25	işin	-1,838	0,005	1,01
altcoine	1,020	0,001	1,06	kripto para piyasalarında	-1,592	0,016	1,01
altcoini	1,232	0,001	1,03	kıdemli	2,083	0,002	1,01
açıklama bu	-1,592	0,001	1,03	matic	-1,381	0,012	1,04
balinaların	1,943	0,000	1,16	merkezi	-1,728	0,016	1,18
bu	0,501	0,000	1,27	milyon dolar	-1,492	0,007	1,05
bu seviyelere	2,395	0,000	1,06	milyon dolarlık	1,156	0,001	1,09
ceosu	1,643	0,013	1,02	nasıl alınır	-1,897	0,004	1,01
dahil	-1,993	0,000	1,02	ne	0,592	0,002	1,26
dakika	1,213	0,000	1,07	nft	-1,680	0,000	1,07
devam	-0,876	0,002	1,06	piyasada son	-1,738	0,002	1,10
diğerleri	-1,269	0,001	1,06	ralli	1,056	0,001	1,04
dolarlık btc	-1,915	0,001	1,05	söylentileri	1,542	0,020	1,01
durum	-0,647	0,000	1,32	sıcak	1,434	0,000	1,02
durum ne	-1,443	0,003	1,39	terra	1,717	0,000	1,18
duyuru	1,196	0,018	1,18	ve fiyat	-1,693	0,019	1,20
ekibinden	-1,905	0,006	1,11	yüzde	1,484	0,000	1,15
elon	1,938	0,000	1,06	yıl	2,525	0,000	1,01
ile	-1,204	0,000	1,11	özel	-2,710	0,000	1,02

Birinci analiz sonucu ortaya çıkartılan 39 kelime ve kelime grubundan, 20 tanesinin okunma oranı üzerinde olumu etkiye sahip olduğu, 19 tanesinin ise olumsuz etkiye sahip olduğu belirlenmiştir. Modelde pozitif en büyük katsayıya sahip ilk üç değişken sırasıyla; "yıl", "bu seviyelere" ve "kıdemli" değişkenleridir. Negatif olarak en büyük katsayıya sahip üç değişken ise şunlardır; "özel", "dahil" ve "dolarlık btc". Bu analizde R^2_{adj} %59,31 olarak hesaplanmıştır, başka bir ifadeyle, oturum sayısındaki değişkenliğin %59,31'i modeldeki 39 bağımsız değişken tarafından açıklanabilmektedir.

Tablo 5'te ise standart MM teknikleriyle ele alınan ikinci analizde elde edilen modele yer verilmiştir. Bu tabloda 30 değişkene ait katsayılar, *p*-değerleri ve VIF değerleri gösterilmektedir. İkinci modelin R^2 değeri %52,49, R^2_{adj} (düzeltilmiş R^2) değeri ise %49,27 olarak hesaplanmıştır.

Tablo 5

İkinci analizin model özeti

Değişkenler	Katsayılar	*p*-değeri	VIF	Değişkenler	Katsayılar	*p*-değeri	VIF
Constant (sabit)	3,8735	0,000		ralli	0,788	0,020	1,06
airdrop	-1,047	0,020	1,13	rekor	-1,485	0,009	1,21
altcoin	0,944	0,000	1,20	satış	2,017	0,000	1,19
al	-1,157	0,015	1,06	sebep	1,799	0,005	1,16
avalanche avax	1,181	0,007	1,07	son durum	-1,350	0,000	1,19
balina	0,917	0,004	1,03	sıcak	1,517	0,000	1,02
bekle	1,147	0,002	1,02	taban	-0,900	0,025	1,04
coinbase	1,402	0,003	1,07	tarih	1,022	0,001	1,04
dakika	0,823	0,001	1,07	terra ekip	2,411	0,002	1,13
devam	-0,851	0,004	1,02	ust	1,384	0,000	1,04
durum mana	1,289	0,022	1,18	uç	2,399	0,000	1,04
düş	1,734	0,001	1,02	vitalik	3,083	0,000	1,00
elon	1,586	0,000	1,01	yıl	1,786	0,000	1,07
inu	1,020	0,001	1,06	çek	-1,441	0,008	1,11
luna	1,465	0,002	1,02	özel	-2,173	0,003	1,00
nft	-1,413	0,000	1,16				

İkinci analiz sonucu ortaya çıkan 30 kelime ve kelime grubundan, 21 tanesinin okunma oranı üzerinde olumu etkiye sahip olduğu, 9 tanesinin ise olumsuz etkiye sahip olduğu belirlenmiştir. Modelde pozitif en büyük katsayıya sahip ilk üç değişken sırasıyla; "vitalik", "terra ekip" ve "uç" değişkenleridir. Negatif olarak en büyük katsayıya sahip üç değişken ise şunlardır; "özel", "rekor" ve "çek". Bu analizde hesaplanan R^2_{adj} %49,27 değeri ile oturum sayısındaki değişkenliğin %49,27'si modeldeki 30 bağımsız değişken tarafından açıklanabilmektedir. Her iki modelin karşılaştırması Tablo 6'da sunulmuştur.

Tablo 6

Model karşılaştırma tablosu

Modeller	R^2_{adj} (düzeltilmiş R^2)	Değişken Sayısı	Güven Aralığı
Birinci Model (farklı yaklaşım)	%59,31	39	%97,5
İkinci Model (standart MM yaklaşımı)	%49,27	30	%97,5

Birinci modelin R^2_{adj} değerinin, ikinci modele göre yaklaşık %10 daha yüksek olduğu tespit edilmiştir. Ayrıca birinci modelde istatistiksel olarak anlamlı değişken sayısı 39 iken, ikinci modelde bu sayı 30'dur.

Sonuç ve Yorum

Birinci ve ikinci analiz sonucu elde edilen %59,31 ve %49,27'lik R^2_{adj} değerleri oldukça değerli çıktıları içerisinde barındırmaktadır. Değişkenlerin metinlerden yaratılarak veri kümesinin oluşturulmasını içeren bu tür bir çalışma için çok yüksek R^2_{adj} değerleri beklemek doğru bir değerlendirme olmayacaktır. Çünkü Facebook gibi sosyal medya mecraları, dışarıya kapalı olan kendi geliştirdikleri algoritmalarla, bazı içerikleri öne çıkartırken bazılarını ise geri plana atmaktadır. Bunu yaparken o anın gündemi, içerikte geçen anahtar kelimeler, içeriğin konusu, takipçi sayıları ve kullanıcı etkileşimleri gibi çok sayıda parametreyi kullanmaktadır. Diğer bir ifadeyle, hangi içeriklerin öne çıkarılacağını veya kullanıcılara daha fazla gösterileceğini algoritmalar belirlemektedir. Bu algoritmalarda yer alan parametrelerin tamamı içerik üreticiler tarafından bilinmediği ve hesaplanamadığı için, bu çalışmanın açıklayamadığı birinci analiz için %40,69 ve ikinci analiz için %50,73'lük geriye kalan kısmın, bunun gibi bilinmeyen parametrelerden kaynaklandığı şeklinde bir değerlendirme yapılmıştır.

Yapılan iki farklı analizin R^2_{adj} sonuçları karşılaştırıldığında, bu çalışmanın ele aldığı probleme özel uygulanan farklı yaklaşımın (~%59), standart MM tekniklerine (~%49) göre yaklaşık %10 daha iyi performans ürettiği görülmektedir. Bunun nedenleri ortaya çıkan sonuçlar göz önüne alınarak açıklanmaya çalışılmıştır. Açıklamaya geçilmeden önce, yapılan iki analizin ayrıştığı noktaları tekrar vurgulamakta fayda görülmektedir. Birinci ve ikinci analiz, çalışmanın yöntem kısmında açıklandığı gibi iki farklı yaklaşımı ele almaktadır. Birinci analiz için veri kümesi, kelime köklerine indirgenmeden (stemming) ve tek başına anlam ifade etmeyen kelimeler (stop-word) çıkarılmadan hazırlanmıştır. İkinci analiz için ise kelimeler köklerine indirgenmiş ve tek başına anlam ifade etmeyen kelimeler çıkarılarak veri kümesi hazırlanmıştır.

Elde edilen sonuçlarda birinci veri kümesi içerisinde yer alıp ikincide yer almayan "bu" kelimesi incelenmiştir. "Bu" kelimesinin; "açıklama bu", "bu" ve "bu seviyelere" şeklinde üç farklı değişken olarak birinci modelde yer aldığı görülmektedir. Tablo 7'de, içerisinde "bu" kelimesi geçen değişkenler ve katsayıları verilmektedir.

Tablo 7

İçerisinde "bu" geçen değişkenler listesi

Değişkenler	Katsayılar	*p*-değeri
açıklama bu	-1,592	0,001
bu	0,501	0,000
bu seviyelere	2,395	0,000

Tablo 7'deki "bu" ifadesi içeren değişkenlere ait katsayılar ele alındığında, "bu" kelimesinin, "açıklama" kelimesiyle beraber "açıklama bu" şeklinde kullanımının, okunma sayısını olumsuz etkilediği gözlemlenmiştir. Öte yandan "bu" kelimesinin, kendinden sonra gelen "seviyelere" kelimesiyle beraber kullanıldığında ise oturum sayısını dramatik bir şekilde arttırdığı

görülmektedir. Bunların yanı sıra "bu seviyelere" kelime grubunun, başlıklarda sadece "bu" kelimesinin kullanılmasına kıyasla, beş kata yakın bir oranda, okunma değişkeni üzerinde olumlu etkisinin olduğu görülmektedir. Eğer birinci analiz için metin işleme ve değişken yaratma aşamalarında, "bu" kelimesi tek başına anlam ifade etmeyen bir kelime olarak etiketlenip veri kümesi dışına alınsaydı, 39 değişken içerisinden en yüksek ikinci olumlu etkiye sahip olarak keşfedilen "bu seviyelere" kelime grubu belirlenemeyecekti. Araştırmaya daha fazla açıklama getirmek ve analiz sonucu ortaya çıkan teorik bilgi ile pratikteki uygulamanın tutarlı bir şekilde örtüşüp örtüşmediğini ortaya koyabilmek için, örnek olarak ifade edilen "bu seviyelere" değişkenine ait ek araştırmalar yapılmıştır. Bu bağlamda, ilgili değişkenin neden okunma sayısını yüksek oranda olumlu etkilediğine dair izahat yapılmaya çalışılmıştır. Öncelikle veri kümesi içerisinde yer alan "bu seviyelere" kelime grubunun kullanıldığı başlıklar incelenmiştir. Bu kapsamda iki başlık değerlendirilmiştir: "tarihi fraktal sinyal veriyor shiba inu shib bu seviyelere ulaşabilir" ve "bitcoin btc ve avax bu seviyelere düşebilir detaylı analiz". Bu iki başlığın, ilgili haberlerde bahsedilen kripto paraların değerleri hakkında öngörüler içeren analizler olduğu dikkat çekmektedir. Hali hazırda "bu seviyelere" kelime grubunun, kendi başına üstü kapalı olarak bir değer belirten ifade olduğu görülmektedir. Ayrıca "bu seviyelere" kelime grubunun başlık içerisindeki konumuna bakıldığında, kilit noktada yer aldığı gözlemlenmektedir. Çünkü "bu seviyelere" kelime grubu, bu iki haberde bahsedilen kripto paraların hangi seviyeye ulaşacağı (yükseleceği) ve düşeceği hakkında (bu kelimelerin hemen öncesinde kullanılarak) gizli bir bilgi barındırmaktadır. Buradan kelime ve kelime gruplarının başlıklarda hangi konumda kullanılacağı, başka bir ifadeyle kelimelerin sırasının dahi önemli olabileceği görülmektedir. Bu konu, bu alanda yapılacak sonraki çalışmalarda ele alınabilir. İki başlığa geri dönülecek olunursa, başlıklarda "bu seviyelere" kelime grubuyla verilen üstü örtülü bilginin, okuyucuda merak uyandırdığı değerlendirilmektedir. Okuyucu, başlıkta üstü kapalı olarak verilen bu bilgiye, ancak içeriğin tamamını okuyarak ulaşabilir. Böylece kullanıcı eğer haberde bahsedilen ilgili konuyla ilgileniyorsa, bu başlığa tıklayıp okumak isteyecektir. Bu çerçevede değerlendirildiğinde, "bu seviyelere" kelime grubunun kullanıcılarda merak uyandıran bir etkiye sahip olduğu ve bu nedenle oturum sayılarını olumlu olarak önemli derecede etkilediği düşünülmektedir.

Tablo 4'te görülebileceği gibi birinci modelin "bu", "açıklama bu" ve "bu seviyelere" kelime ve kelime gruplarına ek olarak, "durum ne", "ile", "ne", "ve fiyat" olmak üzere tek başına anlam ifade etmeyen (stop-word) kelimeler içerdiği görülmektedir. İkinci analiz için standart MM teknikleriyle hazırlanan veri kümesinde ise bu değişkenlerin hiçbiri bulunmamaktadır. Bu değişkenlerin istatistiksel olarak anlamlı olması, birinci modelin ikinci modele göre daha yüksek performansa sahip olmasına katkı sağlamaktadır.

Diğer taraftan, birinci analizde kelimelerin köklerine inilmeden veri kümesi hazırlanmıştı. Aynı örnekle bu konu da açıklanabilmektedir. "Bu seviyelere" kelime grubunda eğer kelime köküne indirgeme işlemi yapılsaydı, değişken sadece "seviye" olarak ele alınabilirdi. Fakat bir önceki paragrafta açıklanmaya çalışılan, "bu seviyelere" kelime grubunun başlık içerisindeki kilit rolü ortaya çıkarılamayacaktı. Ayrıca standart MM teknikleriyle ele alınan ikinci modelde, Tablo 5'te görülebileceği gibi "uç", "al", "bekle", "çek" ve "düş" kelime kökleri istatiksel olarak anlamlı bulunmuştur. Fakat elde edilen bu bilgi, kelimelerin başlıklarda tam olarak nasıl kullanılması gerektiği hakkında bilgi vermemektedir. Oysaki farklı yaklaşımın önerildiği birinci analizde, kelimelerin hangi kullanımlarının etkili olduğu açıkça görülebilmektedir. Başlıklarda kullanılan her

bir kelimenin nasıl kullanıldığı ve bu kelimelerin konumunun dahi önemli olduğu bir problemde, yalnızca köklerden oluşan bir dizi kelime hiçbir anlam ifade etmemektedir. Çünkü kelime köklerini yan yana ekleyerek bir başlık oluşturulamaz ve elde edilen sonuçlar pratikte kullanılamazdı. Başka bir deyişle, bu araştırmanın aradığı cevaplara ulaşılamayacaktı. Daha açık ifade etmek gerekirse, yalnızca köklerden oluşan bir dizi kelime grubu içerisinden yeni başlıklar yaratmak mümkün değildir. İçeriği doğru tanımlayan tam bir başlık yaratmak için, Türkçe dil bilgisine uygun çekim ekleri, yapım ekleri, bağlaçlar ve edatlar gibi bütün bir kullanıma ihtiyaç duyulmaktadır. Özetle, ikinci modelin değişkenlerinden biri olan "uç" kelime köküyle "uçtu", "uçacak", "uçuyor" veya "uçabilir" gibi çok sayıda farklı kelime üretilebilir. Bunlardan hangisinin kullanımının başlıkların okunmasında etkili olabileceği ikinci model ile bilinememektedir.

Tüm bu bilgiler ışığında, teoride ortaya çıkarılan bilginin, pratikte tutarlı bir şekilde örtüştüğü gözlemlenmektedir. Bu tutarlılık durumu ve her iki analiz için hesaplanan R^2_{adj} değerleri, bu çalışma için ele alınan problem özelinde uygulanan farklı yaklaşımın, standart MM tekniklerine göre daha yüksek bir performans sağlayarak geçerli bir yaklaşım olduğunu göstermektedir.

Yapılan araştırmalarda ve literatürde, bu çalışmanın konu edindiği problemi, yaklaşım olarak bu şekilde ele alan bir çalışmaya rastlanmamıştır. Çalışma bu yönüyle özgün bir nitelik taşımaktadır. Benzer MM problemlerinde bu araştırmada uygulanan yaklaşımın kullanılabileceği gösterilmiştir.

Bununla birlikte, araştırma sonuçlarının uygulamada kullanımı hakkında bazı öneriler verilebilir. Bu çalışmada gerçekleştirilen iki farklı analiz içerisinden performansı daha yüksek olan farklı yaklaşımın uygulandığı birinci modelde yer alan kelime ve kelime grupları Şekil 5'te kelime bulutu olarak görselleştirilmiştir.

Şekil 5

Kelime bulutu

Belirlenen kelime ve kelime grupları, ilgili organizasyondaki konu uzmanları tarafından incelenerek değerlendirilebilir. Böylece yeni oluşturulacak içeriklere ışık tutulabilir. Ortaya çıkarılan sonuçlarla, platformun Facebook takipçileri için özel çalışmalar yapılabilir. Ayrıca organizasyon, bu çalışmaya benzer analizleri Twitter veya Telegram gibi farklı sosyal medya

mecraları için de gerçekleştirebilir. Böylece farklı sosyal medya mecraları için farklı içerik stratejiler uygulanarak, mecraya özel içerikler üretilip verimlilik arttırabilir. Ayrıca içeriklerin okunma oranlarına göre bu mecralardaki takipçi kitlesi profillenerek, kitlelere özel içerikler yayınlanabilir. Genel olarak bakıldığında, analiz sonucu elde edilen bilgilerin, organizasyon içerisindeki uzmanlar tarafından değerlendirilerek BY ve MM kavramlarının da amacına uygun olarak, şirkete katma değer yaratacak uygulamalara dönüştürülebileceği görülmektedir. Elbette bu çalışma üç aylık dönemde yayınlanan içerikleri kapsama aldığı için, belirlenen kelime ve kelime grupları, gündemin değişimine göre belirli bir süre sonra geçerliliğini yitirebilir. Fakat belirli dönemlerde bu çalışmada sunulan yöntem ve yaklaşım tekrarlanarak güncel gündeme uygun belirlenecek yeni kelime ve kelime gruplarıyla veri bilimine dayalı sürdürülebilir bir büyüme yakalanabilir.

Sonuç olarak, bu çalışmada organizasyon içerisinde oluşan ham verinin BY alanı içerisinde yer alan MM süreçleriyle işlenerek değerli bilgiye dönüştürülmesi üzerine farklı bir yaklaşım uygulanmıştır. Bu çerçevede iki farklı analiz yapılmış ve sonuçlar karşılaştırılmıştır. Elde edilen sonuçlara göre çalışmanın ele aldığı problem özelinde uygulanan farklı yaklaşımın performansının, standart MM tekniklerine göre daha başarılı olduğu tespit edilmiştir.

İzin ve Katkı Bildirimleri

Etik Kurul İzni

Yazarlar makalede etik kurul iznine ihtiyaç duyulmadığını beyan etmiştir.

Yazarlık Katkısı

Levent Kurt: Fikir/Kavram, Kavramsal arka plan, Metodoloji, Veri toplama, Veri analizi, Veri görselleştirme, Yazım.

Oya Gürdal: Fikir/Kavram, Kavramsal arka plan, Değerlendirme ve inceleme.

*İnci Batmaz:*Metodoloji, Veri analizi, Değerlendirme ve inceleme.

Kaynakça

Aggarwal, C.C. ve Zhai, C. (2012). *Mining Text Data*. Springer. https://doi.org/10.1007/978-1-4614-3223-4_1

Albayrak, A.S. (2006). *Uygulamalı Çok Değişkenli İstatistik Teknikleri*. Ankara: Asil Yayın.

Altınok, G., Karagöz, P., and Batmaz, İ. (2021). Learning to Rank by Using Multivariate Adaptive Regression Splines and Conic Multivariate Adaptive Regression Splines. *Computational Intelligence*, 37, 371-408. doi: 10.1111/coin.12413.

Batmaz, İ. (2021). *Doğrusal Regresyon*. İnci Batmaz'a ait ders notları. Orta Doğu Teknik Üniversitesi İstatistik Bölümü, Ankara.

Batmaz, İ., Danışoğlu, S., Kartal-Koç, E., and C. Yazıcı. (2017). A Data Mining Application to Deposit Pricing: Main Determinants and Prediction Models. *Applied Soft Computing (for Business Analytics)*, 60, 808-819. https://doi.org/10.1016/j.asoc.2017.07.047

Batmaz, İ., Karagöz, P. and G. Serdar. (2017). A Comparative Study on Learning to Rank with Computational Methods. *2017 IEEE International Conference on Big Data (IEEE Big Data 2017)*. Boston, USA. DOI: 10.1109/BigData.2017.8258135

Berson, A., Smith, S.J. ve Thearling, K. (1999). *Building Data Mining Applications for CRM*. New York: McGraw-Hill.

Bifet, A. (2013). Mining Big Data in Real Time. *Informatica*, 37(1), 15-20.

Choi, S., Shin, H. ve Kang, S-S. (2021). Predicting Audience-Rated News Quality: Using Survey, Text Mining, and Neural Network Methods. *Digital Journalism*, *9*(1), 84-105. https://doi.org/10.1080/21670811.2020.1842777

Çakmak, T. ve Eroğlu, Ş. (2020). Sosyal Medyada Kullanıcı Etkileşimi ve İçerik Kategorizasyonu: Ankara'daki Halk Kütüphanelerinin Facebook Gönderilerinin Analizi. *Türk Kütüphaneciliği*, 34(2), 160-186. https://doi.org/10.24146/tk.706882

Dawei, J. (2011). The Application of Data Mining in Knowledge Management. 2011 International Conference on Management of e-Commerce and e-Government, *IEEE Computer Society*, 7-9. https://doi.org/10.1109/ICMeCG.2011.58

Doğan, K. ve Arslantekin, S. (2016). Büyük Veri: Önemi, Yapısı ve Günümüzdeki Durum. *DTCF Dergisi*, 56(1), 15-36. doi: 10.1501/Dtcfder_0000001461

Englmeier, K. (2021). The Role of Text Mining in Mitigating the Threats from Fake News and Misinformation in Times of Corona. *Procedia Computer Science*, *181*, 149–156. https://doi.org/10.1016/j.procs.2021.01.115

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37-54. https://doi.org/10.1609/aimag.v17i3.1230

Jo, T. (2019). *Text Mining: Concepts, Implementation, and Big Data Challenge*. Springer International Publishing. https://doi.org/10.1007/978-3-319-91815-0

Kanji, G. K. (2006). *100 Statistical Tests*. California: SAGE.

Karasar, N. (2022). *Bilimsel Araştırma Yöntemi: Kavramlar İlkeler Teknikler*. Ankara: Nobel.

Larose, D.T. ve Larose, C.D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. (2. Edition). Hoboken, New Jersey: John Wiley & Sons Inc.

Lyu, F. ve Choi, J. (2020). The Forecasting Sales Volume and Satisfaction of Organic Products through Text Mining on Web Customer Reviews. *Sustainability*, *12*, 4383. https://doi.org/10.3390/su12114383

Montgomery, D. C., Peck, E. A. ve Vining, G. G. (2013). *Doğrusal Regresyon Analizine Giriş* (5. Baskı). (M.A. Erar, Çev.) Nobel (2012).

Natarajan, M. (2005). Role of Text Mining in Information Extraction and Information Management. *DESIDOC Bulletin of Information Technology*, 25(4), 31-38. http://dx.doi.org/10.14429/dbit.25.4.3663

Nosakhare, U.H. ve Bright, A.F. (2017). Evaluation of Techniques for Univariate Normality Test Using Monte Carlo Simulation. *American Journal of Theoretical and Applied Statistics*, 6(5-1), 51-61. DOI: 10.11648/j.ajtas.s.2017060501.18

Özdemirci, F. (2018). Sağlık Bilgi Sistemleri Yönetimi ve Toplumsal Bellek/Gelecek Açısından Değerlendirilmesi. *Bilgi Yönetimi Dergisi*, *1*(2), 149-155. https://dergipark.org.tr/tr/pub/by/issue/40526/500294

Pektaş, A. O. (2013). *SPSS İle Veri Madenciliği*. İstanbul: Dikeyeksen.

Savin, N. E. ve White, K. J. (1977). The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors. *Econometrica*, *45*(8), 1989–1996. https://doi.org/10.2307/1914122

Schonlau, M., Guenther, N. ve Sucholutsky, I. (2017). Text Mining with N-Gram Variables. *The Stata Journal*, 17(4), 866-881.

Silwattananusarn, T. ve Tuamsuk, K. (2012). Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 2 (5), 13-24. doi: 10.5121/ijdkp.2012.2502 13

Tan, Z., He, C., Fang, Y., Ge, B. ve Xiao, W. (2018). Title-Based Extraction of News Contents for Text Mining. *IEEE Access*, 6, 64085-64095. DOI: 10.1109/ACCESS.2018.2877592

Trzepiecinski, T., Szpunar, M. ve Kascak, L. (2021) Modeling of Friction Phenomena of Ti-6Al-4V Sheets Based on Backward Elimination Regression and Multi-Layer Artificial Neural Networks. *Materials*, 14, 2570. https://doi.org/10.3390/ma14102570

Tseng, W- T. (2020). Mining Text in Online News Reports of COVID-19 Virus: Key Phrase Extractions and Graphic Modeling. *English Teaching & Learning*, 44, 439-449. https://doi.org/10.1007/s42321-020-00070-2

Uysal, M. ve Günay, S. (2001). Durbin-Watson Ölçütüne Göre Kararsızlık Bölgesinde Bulunan Negatif Otokorelasyon İçin Bazı Testler. *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, 2(2), 277-284. https://earsiv.anadolu.edu.tr/xmlui/handle/11421/802

Wang, H. ve Wang, S. (2008). A Knowledge Management Approach to Data Mining Process for Business Intelligence. *Industrial Management & Data Systems*, 108(5), 622-634. https://doi.org/10.1108/02635570810876750

Wang, Z.H. ve Zou, Z.J. (2018). Quantifying Multicollinearity in Ship Manoeuvring Modeling by Variance Inflation Factor. *In Proceedings of the ASME 2018 37th International Conference on Ocean, Offshore and Arctic Engineering*. https://doi.org/10.1115/OMAE2018-77121

Xiong, S., Lu, S., Shang, F., Li, X., Yan, J. ve Cen, K. (2021). Online Predicting PCDD/F Emission By Formation Pathway İdentification Clustering and Box-Cox Transformation. *Chemosphere*, 274. https://doi.org/10.1016/j.chemosphere.2021.129780