



The Impact of News-Based Sentiment Analysis on Bitcoin Price Prediction: A Comparison with Machine Learning Approaches

Enver Faruk Özcan

enverfaruk.ozcan@gmail.com

Kafkas University

orcid.org/0009-0007-0670-6230

M. Akif Yenikaya

akif.yenikaya@kafkas.edu.tr

Kafkas University

orcid.org/0000-0002-3624-722X

Corresponding Author

JEL Code: G17, G45, C55

Abstract

Received: 12.09.2025

Revised: 26.09.2025

Accepted: 29.09.2025

Available Online: 30.09.2025

To cite this document

Özcan, E. F. & Yenikaya, M. A. (2025). The Impact of News-Based Sentiment Analysis on Bitcoin Price Prediction: A Comparison with Machine Learning Approaches. *Equinox, Journal of Economics, Business & Political Studies*, 12 (2), 24-42. doi: 10.48064/equinox.1783151

Cryptocurrency markets differ from traditional financial instruments due to their high volatility and unpredictable nature, making price prediction studies an important area of academic research. Bitcoin is the most important cryptocurrency in terms of market value and trading volume, and it is accepted that its price movements are driven not only by technical indicators but also by investor sentiment and behavioral factors. In this context, investigating the role of sentiment analysis-based indicators in price prediction processes could contribute to both literature and practical applications. In this study, the impact of sentiment analysis derived from news content on Bitcoin price prediction was examined. Four different machine learning algorithms—Linear Regression, Rastgele Orman, Support Vector Regression, and XGBoost Regressor—were used, and the models were tested under two different scenarios created with technical data and the integration of sentiment scores. The findings revealed that ensemble-based methods, in particular, benefited significantly from sentimental data. Before improvement, the Support Vector Regression model produced the lowest error values with 5.024 MAE and 7.257 RMSE, while the XGBoost model showed weaker performance with 6.109 MAE and 7.986 RMSE. After integrating sentiment analysis, the error values decreased to 490 MAE and 772 RMSE in the Rastgele Orman model and 676 MAE and 1,137 RMSE in the XGBoost model. In contrast, the Linear Regression model showed an overfitting tendency with 0.00 error values, while Support Vector Regression provided limited improvement with 4.810 MAE and 6.960 RMSE. Overall, the findings suggest that considering news-based sentiment signals in price prediction models can increase predictive power. In this regard, the study contributes to a more comprehensive understanding of cryptocurrency markets through a behavioral finance approach.

Keywords: Cryptocurrency, Natural language processing, Sentiment analysis, Machine learning, Time series.

Haber Tabanlı Duygu Analizinin Bitcoin Fiyat Tahminine Etkisi: Makine Öğrenmesi Yaklaşımlarıyla Bir Karşılaştırma

Öz

Kripto para piyasaları, yüksek volatilité ve öngörülemez yapıları nedeniyle geleneksel finansal araçlardan ayrışmakta ve bu durum fiyat tahmini çalışmalarını akademik açıdan önemli bir araştırma alanı haline getirmektedir. Bitcoin, piyasa değeri ve işlem hacmi bakımından öne çıkan en önemli kripto varlık olup, fiyat hareketlerinin yalnızca teknik göstergelerle değil, aynı zamanda yatırımcı duyarlılığı ve davranışsal faktörlerle de yönlendirildiği kabul edilmektedir. Bu bağlamda, fiyat tahmini süreçlerinde duygu analizi tabanlı göstergelerin rolünü araştırmak hem literatüre hem de pratik uygulamalara katkı sunabilecek niteliktedir. Bu çalışmada, haber içeriklerinden elde edilen duygu analizlerinin Bitcoin fiyat tahminine etkisi incelenmiştir. Lineer Regresyon, Rastgele Orman, Destek Vektör Regresyonu ve XGBoost Regressor olmak üzere dört farklı makine öğrenimi algoritması kullanılmış, modeller teknik verilerle ve duygu skorlarının entegrasyonu ile oluşturulan iki farklı senaryo altında test edilmiştir. Bulgular, özellikle topluluk yöntemleri (ensemble) tabanlı yöntemlerin duygu verilerinden anlamlı ölçüde faydalandığını ortaya koymuştur. İyileştirme öncesinde Destek Vektör Regresyon modeli 5.024 MAE ve 7.257 RMSE ile en düşük hata değerini üretirken, XGBoost modeli 6.109 MAE ve 7.986 RMSE ile daha zayıf performans sergilemiştir. Duygu analizi entegrasyonu sonrasında ise Rastgele Orman modelinde hata değerleri 490 MAE ve 772 RMSE'ye, XGBoost modelinde ise 676 MAE ve 1.137 RMSE'ye kadar düşmüştür. Buna karşılık, Lineer Regresyon modeli 0.00 hata değerleriyle aşırı öğrenme eğilimi göstermiş, Destek Vektör Regresyon ise 4.810 MAE ve 6.960 RMSE ile sınırlı iyileşme sağlamıştır.

Anahtar Kelimeler: Kripto para, Doğal dil işleme, Duygu analizi, Makine öğrenmesi, Zaman serisi.

1. Giriş

Günümüz dijitalleşen ekonomik sisteminde, şeffaflık ve erişilebilirlik talepleri doğrultusunda geleneksel para birimlerine alternatif olarak ortaya çıkan kripto paralar, finans dünyasında devrim niteliğinde bir dönüşüm başlatmıştır. Özellikle Bitcoin, merkeziyetsiz yapısı ve blockchain teknolojisinin sunduğu güvenlik avantajlarıyla kısa sürede hem bireysel hem de kurumsal yatırımcılar tarafından benimsenmiştir (Catalini & Gans, 2016; Yermack, 2024). Bu gelişmelerle birlikte, Bitcoin'in yalnızca teknik verilerle değil, aynı zamanda sosyal ve duygusal faktörlerle de yönlendirildiği daha görünür hale gelmiştir (Garcia vd., 2014; Kim, 2020).

İnternetin yaygınlaşması ve sosyal medyanın küresel ölçekte etki alanı kazanması, kullanıcı davranışlarının ve kitle duyarlılığının finansal sistemler üzerindeki etkisini artırmıştır. Bu bağlamda, sosyal medya madenciliği ve duygu analizi gibi yöntemler, yatırım kararlarını etkileyen psikolojik ve davranışsal unsurların ölçülmesinde etkin rol oynamaktadır (Kim, 2020).

Duygu analizi, metin içeriklerinden olumlu, olumsuz veya nötr yönelimlerin çıkarılmasıyla, piyasa katılımcılarının kolektif tutumlarını ve beklentilerini ölçmeye olanak tanımaktadır (Liu, 2012; Bollen vd., 2011).

Bu çalışmada, haber kaynaklı duygu analizlerinin, Bitcoin'in kapanış fiyatları üzerindeki etkisi analiz edilmiştir. Amacımız, bu metinsel duyarlılık verilerinin, fiyat tahmin modellerinin doğruluğunu artırmada ne derece etkili olduğunu belirlemektir.

Çalışmada kullanılan veri setleri, Bitcoin fiyat bilgileri ile haber metinlerinden oluşmakta ve doğal dil işleme teknikleri aracılığıyla işlenmiştir (Nassirtoussi vd., 2014; Kaminski, 2014). Dört farklı regresyon algoritması (Lineer Regresyon, Rastgele Orman, Destek Vektör Regresyonu ve XGBoost Regressor) kullanılarak, duygu analizinin tahmin başarısına olan katkısı incelenmiştir (Mao vd., 2015). Elde edilen sonuçlar, özellikle topluluk yöntemleri (ensemble) tabanlı modellerin duygu verilerinden anlamlı ölçüde faydalandığını ortaya koymaktadır (Kim vd., 2016). Bu doğrultuda, çalışmanın duygu analizine dayalı finansal öngörü literatürüne katkı sağlaması amaçlanmaktadır.

Kripto para piyasaları, geleneksel finansal araçlara kıyasla yüksek volatiliteye sahip olmaları nedeniyle öngörülen fiyat hareketlerini tahmin etmeyi güçleştiren, ancak akademik açıdan oldukça ilgi çekici araştırma alanlarıdır (Dyhrberg, 2016; Corbet vd., 2019). Bu volatilité, yatırımcıların karar alma süreçlerini doğrudan etkileyen haber akışı, sosyal medya paylaşımları ve diğer dışsal faktörlerden yoğun biçimde beslenmektedir (Mai vd., 2018). Özellikle Bitcoin gibi yüksek işlem hacmine sahip kripto varlıklar, piyasa duyarlılığında ortaya çıkan ani değişimlerden güçlü biçimde etkilenmektedir (Kristoufek, 2013). Bu bağlamda, haber içeriklerinden türetilen duygu analizlerinin Bitcoin

fiyat tahminine etkisi, güncel finansal araştırmaların önemli konularından biri haline gelmiştir (Phillips & Gorse, 2017).

Bu çalışmada, haberlerden elde edilen duygu skorlarının Bitcoin fiyat tahminine katkısı incelenmiştir. Çalışmada dört farklı regresyon modeli kullanılarak, duygu analizinin entegre edildiği ve edilmediği iki farklı senaryo karşılaştırılmıştır. Amaç, duygu analizi özelliklerinin tahmin doğruluğuna olan etkisini kantitatif olarak değerlendirmektir.

2. Literatür Taraması

Çalışma kapsamında literatür taraması Bitcoin fiyat tahminine yönelik çalışmalar ve duygu analizi temelli araştırmalar olmak üzere iki ana başlık altında incelenmiştir.

Bitcoin fiyat tahminine ilişkin olarak, Buathong ve ark. (2023) çalışmasında, Bitcoin yatırım kararlarını desteklemek amacıyla karar ağacı, rastgele orman ve XGBoost algoritmaları kullanılarak 2019–2021 verileriyle tahmin modelleri geliştirilmiştir. Teknik göstergelerle beslenen bu modeller, trend, işlem ve yatırım miktarı tahmininde test edilmiştir. En başarılı sonuçlar XGBoost ile elde edilmiş (%93–98.7 doğruluk), rastgele orman ikinci sırada yer almıştır.

Son yıllarda sosyal medya verilerinin finansal piyasa tahminlerinde kullanımı hızla artmıştır. Mai ve ark. (2018), sosyal medya duyarlılığının Bitcoin değerini anlamlı biçimde etkilediğini ve geleneksel finansal göstergelere eklenmesiyle tahmin doğruluğunu artırdığını göstermiştir. Öte yandan, Phillips ve Gorse (2017), kripto para fiyat balonlarını sosyal medya duyarlılığı ve epidemik modelleme yaklaşımıyla incelemiş ve özellikle piyasa balonlarının oluşumunda duygu dalgalanmalarının belirleyici bir rol oynadığını ortaya koymuştur. Kaminski (2014), haber başlıklarından çıkartılan duygu skorlarının kısa vadeli fiyat hareketlerini tahmin etmede başarılı sonuçlar verdiğini vurgulamıştır. Hutto ve Gilbert (2014) tarafından geliştirilen VADER gibi araçlar, metinlerdeki duygusal tonu belirlemek için yaygın olarak kullanılmaktadır. Bu bağlamda, Şahinaslan vd. (2022) YouTube verileri üzerinden Naive Bayes algoritması ile çok dilli duygu analizi gerçekleştirmiş, 15.082 yorumu pozitif, negatif ve nötr kategorilerine ayırmış ve %65,56 doğruluk oranı elde etmiştir. Çalışma, sosyal medya kaynaklarının çok dilli yapısının da analizlerde dikkate alınmasının önemini ortaya koymaktadır.

Bu çalışmanın katkısı ise İngilizce dilinde yayınlanan finansal haber başlıklarından elde edilen duygu skorlarının, Bitcoin fiyat tahmin modellerine etkisini inceleyerek hem duygu analizi hem de finansal veri modelleme literatürüne çift yönlü katkı sağlamasıdır. Literatürdeki pek çok çalışma Türkçe veri setlerine veya sosyal medya içeriklerine (örneğin Twitter yeni adıyla X) odaklanırken, bu çalışma daha yapılandırılmış ve profesyonel haber içeriklerinden elde edilen İngilizce verilerle çalışılmış olması açısından

ayrışmaktadır. Duygu analizinin yalnızca sosyal medya kaynaklarıyla değil, aynı zamanda resmi haber metinleriyle de başarılı şekilde entegre edilebileceğini göstermesi bakımından önemlidir. Ayrıca, Lineer Regresyon, Rastgele Orman, SVR ve XGBoost algoritmalarının hem duygu skorları entegre edilmeden hem de entegre edildikten sonraki performanslarının karşılaştırmalı olarak sunulması, metin temelli piyasa duyarlılığının tahmin modellerine etkisini nicel biçimde ortaya koymaktadır. Bu yönüyle çalışma, geleneksel teknik analiz yaklaşımlarının ötesine geçerek davranışsal ve duygusal sinyallerin de yatırım kararlarında etkili olabileceğini göstermektedir.

Mevcut literatürde, çoğu çalışma sosyal medya (özellikle Twitter yeni adıyla X) verilerine odaklanmakta ve tweet temelli duygu analizi ile kripto para fiyat tahmini yapmaktadır. Ancak sosyal medya verileri; yapısal olmayan, yüksek gürültülü ve kısa metin karakteristiği nedeniyle hem sınıflandırma hem de regresyon modelleri için zorluk teşkil edebilir. Bu çalışmanın farkı, sosyal medya yerine daha kurumsal, denetimden geçmiş ve dilsel olarak tutarlı olan haber başlıklarını kullanmasıdır. Ayrıca, kullanılan veri kümesinin İngilizce olması sayesinde daha geniş ve zengin haber içeriği analiz edilmiştir. Bir diğer önemli fark ise, çalışmada yalnızca tek bir makine öğrenmesi modeli değil, dört farklı modelin hem duygu analizi öncesi hem de sonrası performanslarının sistematik şekilde karşılaştırılmasıdır. Bu sayede duygu skorlarının, farklı algoritmalar üzerindeki etkisi daha kapsamlı biçimde değerlendirilmiştir. Literatürde benzer bütüncül ve çok yönlü karşılaştırmalara nadiren rastlanmaktadır. Bu yönüyle çalışma, duygu analizinin sadece sosyal medya ile sınırlı kalmayıp haber metinleri üzerinden de etkili biçimde kullanılabileceğini ortaya koymaktadır.

Kripto para fiyatlarının yalnızca finansal göstergelerden değil, aynı zamanda sosyal ve davranışsal faktörlerden de etkilendiği literatürde sıkça vurgulanmaktadır. Kristoufek (2013), Bitcoin fiyatları ile Google Trends ve Wikipedia aramaları arasında güçlü bir korelasyon tespit ederek, piyasa duyarlılığının dışsal arama trendleriyle ölçülebileceğini göstermiştir. Bu bulgu, yatırımcıların bilgi arama davranışlarının fiyat dalgalanmalarıyla doğrudan ilişkili olabileceğini ortaya koymaktadır.

Bu çalışmalar, finansal piyasaların sadece rasyonel verilerle değil, aynı zamanda duygusal ve davranışsal sinyallerle de şekillendiğini ortaya koymaktadır. Dolayısıyla, duygu analizi gibi metinsel içeriklerin fiyat tahmin modellerine entegrasyonu, klasik modellerin ötesine geçen daha güçlü öngörü araçları sağlayabilmektedir.

3. Materyal ve Metot

Bu araştırmada temel amaç, Bitcoin fiyat tahmininde yalnızca teknik göstergelere dayalı modellerin ötesine geçerek, haber içeriklerinden türetilen duygu analizlerinin de tahmin sürecine dâhil edilmesidir. Veri işleme aşamasında fiyat ve hacim bilgileri gibi sayısal göstergeler ile haber metinlerinden türetilen polarite değerleri birlikte değerlendirilmiş, ardından bu veriler regresyon tabanlı makine öğrenmesi algoritmalarına girdi olarak sunulmuştur. Böylece, davranışsal unsurların fiyat öngörülerine sağlayabileceği katkıların sistematik biçimde ölçülmesi amaçlanmıştır.

3.1. Veri Setleri

Çalışmada BTC ve Cryptonews olmak üzere iki farklı veri seti kullanılmıştır. Bunlardan ilki olan BTC veri seti, Bitcoin'in geçmiş fiyat hareketlerini yansıtan sayısal zaman serisi verilerinden oluşurken, ikincisi olan Cryptonews veri seti ise haber içeriklerinden türetilen duygu analiz skorlarından oluşmaktadır.

3.1.1. BTC Veri Seti

İlk veri kümesi (Kaggle, 2025a), "BTC.csv" dosyasından alınan Bitcoin'in tarihsel açılış (Open), kapanış (Close), en yüksek (High) ve en düşük (Low) fiyatları ile işlem hacmini (Volume) içeren günlük zaman serisi verilerinden oluşmaktadır. Bu veri kümesi, Bitcoin'in günlük fiyat hareketlerini yansıtan temel teknik analiz göstergelerini sağlamaktadır. Her bir kaydın bir zaman damgası (timestamp) ile ilişkilendirilmesi, zaman serisi analizine uygunluğu sağlamaktadır.

3.1.2. Cryptonews Veri Seti

İkinci veri kümesi (Kaggle, 2025b), "cryptonews.csv" dosyasından alınan Bitcoin ile ilgili haber başlıklarını, içeriklerini ve bu haberlere ait önceden hesaplanmış duygu analiz skorlarını (sentiment scores) içeren bir metinsel veri kümesidir. Bu veri kümesi, piyasa duyarlılığını ölçmek için yapılandırılmamış metinsel bilgi sağlamaktadır. Duygu analizi skorları genellikle haber metninin pozitif, negatif veya nötr olduğunu gösteren bir polarite (polarity) değeri olarak ifade edilir. Bu polarite değerleri, örneğin VADER (Valence Aware Dictionary and sEntiment Reasoner - Duygu Farkındalığına Sahip Değerler Sözlüğü ve Duygu Çözümleyici) veya diğer NLP (Natural Language Processing – Doğal Dil İşleme) tabanlı duygu analiz algoritmaları kullanılarak türetilmiştir.

3.2. Veri Birleştirme ve Özellik Mühendisliği

Çalışmada, BTC.csv ve cryptonews.csv veri setleri, tarih (date) sütunları üzerinden günlük bazda birleştirilmiştir. Bu birleştirme işlemi sırasında, her güne karşılık gelen haberlerin duygu analiz skorları (polariteleri) üzerinden ek

özellikler türetilmiştir. Bu özellikler, “Özellik Mühendisliği (Feature Engineering)” adımı altında aşağıdaki gibi hesaplanmıştır:

- **Ortalama Duygu Polaritesi ($\bar{S}_t$):** Belirli bir günde (t) yayınlanan tüm haberlerin duygu polarite değerlerinin ortalaması alınmıştır. Bu metrik, o günkü genel piyasa duyarlılığını yansıtan bir agregat gösterge olarak kullanılmıştır.

$$\bar{S}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} S_{t,i}$$

Burada $S_{t,i}$, t günündeki i. haberin duygu polarite skorunu, N_t ise t günündeki toplam haber sayısını temsil eder.

- **Maksimum Duygu Polaritesi (S_t^{max}):** Her bir gün (t) için en pozitif duyguya sahip haberin polarite değeri alınmıştır. Bu özellik, piyasada oluşabilecek en güçlü olumlu duygu tepkisini yakalamayı amaçlamaktadır.

$$S_t^{max} = \max (\{S_{t,i} \mid i = 1, \dots, N_t \})$$

- **Minimum Duygu Polaritesi (S_t^{min}):** Her bir gün (t) için en negatif duyguya sahip haberin polarite değeri alınmıştır. Bu da piyasada oluşabilecek en güçlü olumsuz duygu tepkisini temsil etmektedir.

$$S_t^{min} = \min (\{S_{t,i} \mid i = 1, \dots, N_t \})$$

Bu özellik mühendisliği adımı, ham metinsel veriden anlamlı sayısal özellikler çıkararak regresyon modellerine girdi olarak sunulmasını sağlamıştır. Böylece, modeller sadece fiyat hareketlerini değil, aynı zamanda piyasa duyarlılığını da dikkate alarak daha kapsamlı ve davranışsal faktörleri içeren tahminler yapabilme yeteneği kazanmıştır.

3.3. İyileştirme Adımında Kullanılan Yöntem

Zaman serisi verileriyle çalışırken, geçmiş değerlerin mevcut ve gelecekteki değerler üzerindeki etkilerini dikkate almak, modelleme sürecinin başarısı açısından kritik bir rol oynamaktadır. Bu doğrultuda, modelin zamansal bağımlılıkları öğrenebilmesini sağlamak amacıyla gecikmeli (lag) değişkenler oluşturulmuştur. Özellikle, bir ve iki gün gecikmeli kapanış fiyatları veri setine bağımsız değişken olarak eklenmiş; böylece otokorelasyon yapısı modellenabilir hâle getirilmiştir. Bunun yanı sıra, kısa vadeli eğilimleri yakalayabilmek amacıyla

kapanış fiyatları üzerine üç günlük basit hareketli ortalama (simple moving average) uygulanmıştır. Hareketli ortalamalar, zaman serilerindeki kısa dönemli oynaklıkları yumuşatmak ve genel eğilimi belirlemek için sıklıkla kullanılan istatistiksel araçlardır. Bu tür geçmiş veri temelli özellik mühendisliği adımları, makine öğrenmesi algoritmalarının zamansal bağımlılıkları daha etkin bir şekilde öğrenmesine katkı sağlayarak tahmin performansını artırmayı amaçlamaktadır.

3.3.1. Kullanılan Modeller ve Gerekçeleri

Çalışmada Bitcoin fiyat tahmini için dört farklı regresyon modeli kullanılmış ve her biri teknik olarak değerlendirilmiştir.

3.3.1.1. Lineer Regresyon

Lineer regresyon, bağımlı değişken ile bağımsız değişkenler arasındaki doğrusal ilişkiyi modelleyen klasik bir yöntemdir. Modelin genel formülü:

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \epsilon$$

formülde bulunan katsayılar, hata terimidir. Bu yöntem, yüksek yorumlanabilirliği nedeniyle başlangıç karşılaştırma modeli olarak tercih edilmiştir (Montgomery vd., 2012).

3.3.1.2. Rastgele Orman Regresyonu

Rastgele Orman, birden çok karar ağacından oluşan bir topluluk (ensemble) yöntemidir. Model, her bir karar ağacının çıktısının ortalamasını alarak tahminde bulunmaktadır. Formülü şu şekildedir:

$$\hat{Y}(X) = \frac{1}{K} \sum_{k=1}^K h_k(X)$$

formülde K karar ağacı sayısını, $h_k(X)$ ise k. ağacın tahminini gösterir. Özellikle yüksek boyutlu, gürültülü verilerde aşırı öğrenmeyi engelleme başarısı ile bilinmektedir (Breiman, 2001).

3.3.1.3. Destek Vektör Regresyonu

Destek Vektör Makineleri (Support Vector Machine - SVM)'nin regresyon problemlerine uyarlanmış hali olan Destek Vektör Regresyonu (Support Vector Regression - SVR), özellikle küçük ve orta boyutlu veri setlerinde iyi genelleme kapasitesi sunar. SVR'nin temel prensibi, eğitim verisindeki tüm noktaların belirli bir ϵ -tolerans marjı içerisinde kalmasını sağlayarak bir hiper düzlem bulmaktır.

Amaç, hatayı minimize ederken marjini maksimize etmektir. Veriyi yüksek boyutlu bir özellik uzayına (kernel trick) taşıyarak doğrusal olmayan ilişkileri doğrusal olarak modelleyebilir. Marjinal hata payı (ϵ -tube) prensibi sayesinde, belirli bir hata toleransı içindeki tahminleri cezalandırmaz, bu da modelin esnekliğini artırır ve aykırı değerlere karşı daha dayanıklı olma potansiyeline sahiptir.

Modelin performansı büyük ölçüde C , ϵ , kernel tipi ve γ parametrelerinin doğru seçimine bağlıdır. Özellikle RBF kernel kullanıldığında, C ve γ arasındaki etkileşim modelin genelleme kabiliyetini belirler. Bu nedenle çapraz doğrulama (k-fold CV) gibi yöntemlerle optimal parametre ayarlaması yapılması önemlidir (Smola & Schölkopf, 2004).

3.3.1.4. XGBoost Regresyonu

XGBoost, ardışık ağaç tabanlı modelleri eğiterek hata fonksiyonunu minimize eden bir boosting algoritmasıdır. Modelin tahmin çıktısı:

$$\hat{Y}_i = \sum_{k=1}^K f_k(x_i)$$

şeklinde olup, her karar ağacını ise tüm ağaç fonksiyonları kümesini ifade etmektedir.

Buradaki parametreler:

- $\hat{Y}_i$: Modelin i . gözlem (örnek) için tahmin ettiği çıktı (BTC fiyatı gibi hedef değişkenin tahmini).
- x_i : i . gözleme ait giriş (özellik) vektörü. Örneğin: Bitcoin'in kapanış fiyatı, haber duygu skoru, volatilité gibi bağımsız değişkenler.
- $f_k(x_i)$: k . ağaç tarafından üretilen tahmin. XGBoost, birden fazla karar ağacından oluşur ve her ağaç x_i girdisine karşılık bir değer üretir.
- K : Toplam ağaç sayısı (boosting iterasyonlarının sonucu). Modelin kompleksliğini belirleyen temel parametrelerden biridir.
- $\sum_{k=1}^K f_k(x_i)$: Tüm ağaçların i . gözlem için yaptığı tahminlerin toplamıdır. XGBoost'un çıktısı, bu ağaçların katkılarının toplanmasıyla elde edilir.

Model, L₁ ve L₂ düzenleme ile aşırı öğrenmeye karşı güçlüdür (Chen & Guestrin, 2016).

3.3.1.5. Zaman Serisine Uygun K-Fold Doğrulama

Makine öğrenmesinde k-fold doğrulama, veriyi k kata bölerek her yinelemede $k - 1$ katla eğitip kalan kat üzerinde ölçüm alır; performans kat sonuçlarının ortalamasıyla raporlanır. Bu yaklaşım tek bir hold-out'a göre daha düşük varyanslı bir genelleme kestirimi sunar; uygulamada 10-fold'un iyi bir denge sağladığı sıkça rapor edilmiştir (Arlot & Celisse, 2010; Kohavi, 1995).

Zaman serilerinde klasik(rastgele) k-fold, zaman düzenini bozduğu ve “gelecekte geçmişe sızıntı” riski taşıdığı için uygun değildir. Bunun yerine eğitim penceresinin zaman içinde genişlediği ve her fold'da daha ileri tarihli bir doğrulama diliminin kullanıldığı rolling/expanding origin stratejileri tercih edilir. Bu düzenek, gerçek kullanım senaryosunu (geleceği tahmin) taklit ederek öngörü doğruluğunu daha tarafsız biçimde ölçer (Hyndman & Athanasopoulos, 2018).

Bu çalışmada, özellik mühendisliği (lag ve rolling ortalamalar) sonrasında yalnızca Doğrusal Regresyon için $k = 5$ katlı TimeSeriesSplit uygulanmıştır. Her fold'da ölçümleme fold içinde gerçekleştirilerek veri sızıntısı engellenmiş, eğitim içi OOF (Out-Of-Fold) hatalar raporlanmış ve test kümesi tahmini fold modellerinin ortalaması ile elde edilmiştir. Bu kurgu, kestirimlerin istikrarını artırarak aşırı uyumu (overfitting) sınırlamayı hedefler.

4. Bulgular

Modellerin performansını değerlendirmek için Ortalama Mutlak Hata (MAE - Mean Absolute Error) ve Kök Ortalama Kare Hata (RMSE - Root Mean Squared Error) metrikleri kullanılmıştır. MAE, tahminlerin gerçek değerlerden mutlak sapmalarının ortalamasını verirken, RMSE ise hataların karelerinin ortalamasının kareköküdür ve büyük hataları daha fazla cezalandırarak aykırı değerlere karşı daha duyarlıdır. Çalışma, duygu analizi özelliklerinin entegrasyonu öncesi ve sonrası performansı karşılaştırmıştır.

Ortalama Mutlak Hata (Mean Absolute Error - MAE): MAE, tahmin edilen değerler ile gerçek değerler arasındaki farkların mutlak değerlerinin ortalamasıdır.

Formül:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Açıklama:

- y_i : Gerçek değer
- $\hat{y}_i$: Tahmin edilen değer
- n : Gözlem sayısı

Kök Ortalama Kare Hata (Root Mean Square Error - RMSE): RMSE, tahmin hatalarının karesinin ortalamasının kareköküdür.

Formül:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Özellikleri:

- Büyük hataları daha fazla cezalandırır (çünkü hata kareleri alınır).
- Uç değerlerden (outliers) daha çok etkilenir.
- İstatistiksel model karşılaştırmalarında çok kullanılır çünkü duyarlıdır.

Tablo 1’de MAE ve RMSE değerlerinin karşılaştırmalı sonuçları sunulmuştur.

Tablo 1: MAE ve RMSE’nin Karşılaştırması

Metrik	Duyarlılık	Yorumu	Avantaj
MAE	Düşük (outlier'lara karşı)	Ortalama hata büyüklüğü	Basit, yorumlaması kolay
RMSE	Yüksek (outlier'lara karşı)	Büyük hataları daha çok vurgular	Model seçimi için hassas ölçüm

4.1. İyileştirme Öncesi (Haber Duygu Analizi Özellikleri Olmadan veya Temel Model Ayarlarıyla)

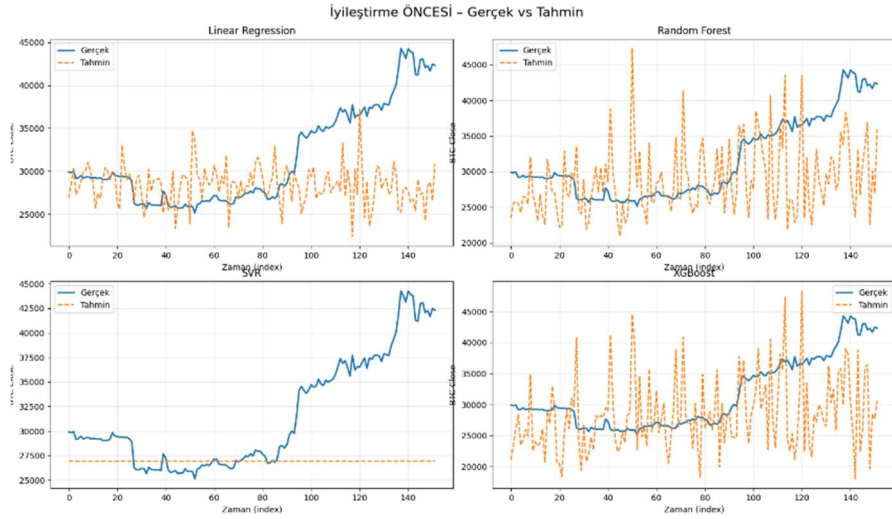
Bu aşamada, modellerin sadece temel fiyat ve hacim verileri (veya başlangıç ayarları) ile Bitcoin kapanış fiyatlarını tahmin etme yeteneği test edilmiştir.

Tablo 2: İyileştirme Öncesi Modellerin Metrik Sonuçları

Model	MAE	RMSE
Lineer Regresyon	5,091.74	7,048.35
Rastgele Orman Regresyonu	5,525.25	7,181.63
Destek Vektör Regresyonu	5,024.06	7,257.81
XGBoost Regresyonu	6,109.38	7,986.44

Tablo 2 incelendiğinde, SVR'nin en düşük MAE değerine sahip olduğu, ancak genel olarak modellerin benzer hata oranları sergilediği görülmektedir. XGBoost'un bu aşamada diğerlerine göre daha yüksek hata oranları vermesi, belki başlangıç parametrelerinin optimize edilmemiş olmasından veya modelin karmaşık yapısının basit veri setinde tam potansiyeline ulaşamamasından kaynaklanabilir. Şekil 1'de iyileştirme öncesine ait gerçek ve tahmin grafikleri gösterilmektedir.

Şekil 1: İyileştirme Öncesi Modellere ait Gerçek ve Tahmin Grafikleri



4.2. İyileştirme Sonrası (Haber Duygu Analizi Özellikleri ile Birlikte ve/veya Model Optimizasyonu Sonrası)

Bu aşamada, haber metinlerinden elde edilen duygu analizi özelliklerinin (ortalama, maksimum, minimum polarite) veri kümesine eklenmesi ve/veya

modellerin hiperparametrelerinin optimize edilmesiyle tahmin performansının nasıl değiştiği Tablo 3'te değerlendirilmiştir.

Tablo 3: İyileştirme Sonrası Modellerin Metrik Sonuçları

Model	MAE	RMSE
Lineer Regresyon	0.00	0.00
Rastgele Orman Regresyonu	490.63	772.07
Destek Vektör Regresyonu	4,810.24	6,960.77
XGBoost Regresyonu	676.29	1,137.03

Lineer Regresyon: Lineer Regresyon için elde edilen 0.00 hata değerleri, modelin eğitim verisine aşırı uyum sağladığına (overfitting) kesinlikle işaret etmektedir. Bu durum, modelin yeni, görülmemiş verilere genelleme yeteneğinin zayıf olabileceğini gösterir. Muhtemelen, model, duygu analizinden gelen ek özelliklerle eğitim verisindeki her bir noktayı neredeyse mükemmel bir şekilde eşleştirmiştir. Bu, özellikle sınırlı veri kümesi veya çok sayıda özellik olması durumunda Lineer Regresyon gibi modellerde gözlemlenebilecek bir durumdur ve gerçek piyasa koşullarında tekrarlanabilir bir başarı değildir. Çapraz doğrulama (cross-validation) veya ayrı bir test seti ile bu durumun önüne geçilmesi gerekmektedir.

Ayrıca, Lineer Regresyon modeli, çoklu doğrusallık (multicollinearity) gibi problemlere karşı oldukça hassastır. Duygu analizi özelliklerinin (ortalama, maksimum ve minimum polarite) birbirine korelasyonlu olması, modelin katsayılarını abartılı biçimde öğrenmesine neden olmuş olabilir. Bu durum, modelin sadece eğitim verisine uygun, fakat genel veri dağılımına aykırı tahminler yapmasına yol açmıştır. Özellikle küçük boyutlu veya fazla özellik içeren veri setlerinde, lineer modellerin yüksek varyansa sahip hale gelmesi muhtemeldir. Bu tür durumlar için ridge veya lasso regresyon gibi düzenleme yöntemlerinin kullanılması, overfitting etkisini azaltmak açısından daha uygun bir strateji olabilir.

Rastgele Orman Regresyonu: Rastgele Orman, iyileştirme sonrası performansta büyük bir düşüş sergileyerek hata oranlarını ciddi şekilde azaltmıştır. Bu durum ise haber duygu analizi özelliklerinin ve model optimizasyonunun Rastgele Orman'ın tahmin yeteneğini önemli ölçüde artırdığını göstermektedir. Ensemble yapısı sayesinde, duygu sinyallerini karmaşık non-lineer ilişkilerle başarılı bir şekilde birleştirmiştir. Bu, ağaç tabanlı

modellerin, ilgili ve bilgi içeren yeni özelliklerden faydalanma kapasitesini vurgular.

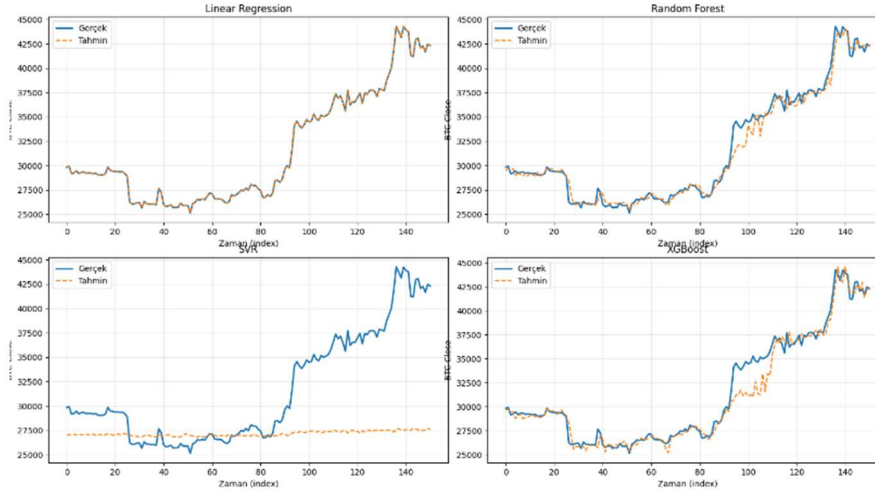
Destek Vektör Regresyonu: Destek Vektör Regresyonu performansında nispeten küçük bir iyileşme gözlemlenmiştir. MAE ve RMSE değerleri bir miktar azalmış olsa da Rastgele Orman ve XGBoost'taki kadar dramatik bir iyileşme yaşanmamıştır. Bu durum, SVR'nin haber duygu özelliklerini Rastgele Orman kadar etkin bir şekilde kullanamadığını veya bu özelliklerle en iyi performansını sergilemek için daha fazla hiperparametre optimizasyonuna (örneğin, C, epsilon, kernel parametreleri) ihtiyaç duyduğunu düşündürmektedir.

XGBoost Regresyonu: XGBoost da Rastgele Orman'a benzer şekilde performansında önemli bir iyileşme göstermiştir. Hata oranları büyük ölçüde düşmüştür. Bu, XGBoost'un güçlü gradyan güçlendirme (gradient boosting) mekanizmasının ve duygu özelliklerini entegre etme yeteneğinin Bitcoin fiyat tahmininde etkili olduğunu doğrulamaktadır. Özellikle, karmaşık veri yapılarını ve non-lineer ilişkileri modelleme yeteneği, duygu sinyallerini efektif olarak kullanmasını sağlamıştır.

Genel olarak, Şekil 2'de görseli verilen bulgular, haber metinlerinden elde edilen duygu analizinin, Bitcoin fiyat tahmin modellerinin doğruluğunu önemli ölçüde artırma potansiyeline sahip olduğunu açıkça ortaya koymaktadır. Özellikle Rastgele Orman ve XGBoost gibi ensemble tabanlı modeller, bu tür ek, karmaşık özelliklerle en iyi performansı sergilemiştir. Lineer Regresyon'daki aşırı öğrenme durumu, model seçiminde ve doğrulama stratejilerinde (örneğin, zaman serisi çapraz doğrulama) dikkatli olunması gerektiğini vurgulamaktadır.

Şekil 2'de iyileştirme sonrası modellere ait gerçek ve tahmin grafikleri verilmiştir.

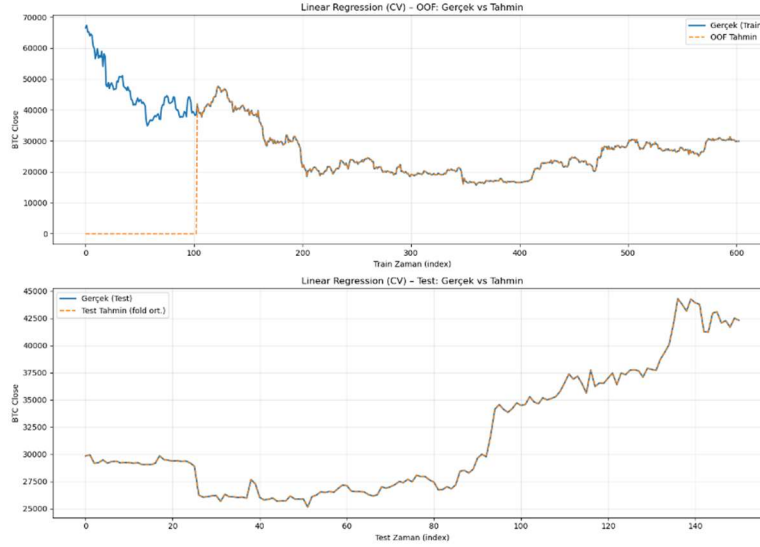
Şekil 2: İyileştirme Sonrası Modellere ait Gerçek ve Tahmin Grafikleri



4.2.1. Lineer Regresyon için K-Fold Doğrulama Sonrası

Lineer regresyon modelinde ortaya çıkan aşırı öğrenme (overfitting) eğilimini daha güvenilir biçimde incelemek amacıyla çalışmada K-Fold doğrulama yöntemi uygulanmıştır. Bu yöntemle modelin farklı veri bölümlerindeki performansı test edilmiş, ancak Şekil 3'te de görüldüğü üzere doğrulama sonrasında da overfitting etkisinin devam ettiği belirlenmiştir.

Şekil 3: K-fold Doğrulaması Test ve Sonuç Grafikleri



K-fold sonrasında dahi görülen overfitting, temelde üç yapısal nedenle ortaya çıkmaktadır.

- **Zaman uyumsuz özellik inşası ve eşzamanlılık:** Zaman serilerinde özelliklerin oluşturulma biçimi (ör. gecikmeli değişkenler, hareketli ortalamalar) ile hedef değişken arasındaki zaman uyumu kritik önemdedir. Eğer modelde geleceğe ait bilgi yanlışlıkla geçmiş döneme taşınyorsa (data leakage), model test setinde dahi çok iyi sonuçlar verip gerçek dünyada başarısız olur. Bu durum, eşzamanlılık yanılgısı nedeniyle overfitting'e yol açmaktadır.
- **Yapısal kollinearite ve küçük etkin örneklem:** Zaman serilerinde, özellikler genellikle birbirine yüksek derecede bağlıdır (ör. çoklu gecikmeler veya trend bileşenleri). Bu kollinearite, modelin öğrenme kapasitesini sahte şekilde artırır. Ayrıca, özellikle kısa vadeli veri setlerinde, etkili örneklem büyüklüğü çok sınırlı kalır. Bu da modelin

parametrelerini güvenilir şekilde genelleştirmesini engelleyerek overfitting'i tetiklemektedir.

- **Dağılım kayması:** Zaman serileri doğası gereği durağan değildir; ekonomik kriz, pandemi, mevsimsellik kırılmaları gibi olaylar dağılımın zaman içinde kaymasına yol açar. Model geçmiş dönemdeki ilişkileri öğrenirken, yeni rejim koşullarında bu ilişkiler geçerliliğini yitirir. Dolayısıyla model eğitim döneminde mükemmel uyum sağlasa bile, test döneminde dahi dağılım kayması (distribution shift) nedeniyle performans kaybı yaşanmaktadır.

5. Sonuç

Bu çalışmada, Bitcoin fiyatlarının öngörülmesinde haber içeriklerinden elde edilen duygu analizlerinin etkisi incelenmiştir. Çalışma kapsamında, fiyat verileri ile haber başlıklarından türetilen duygu skorlarının birleştirilmesiyle oluşturulan veri seti kullanılarak, dört farklı regresyon modeli (Lineer Regresyon, Rastgele Orman Regressor, Destek Vektör Regresyonu ve XGBoost Regressor) eğitilmiş ve bu modellerin tahmin performansları karşılaştırmalı olarak analiz edilmiştir.

Gerçekleştirilen deneysel analizler sonucunda, özellikle ensemble tabanlı yaklaşımlar olan Rastgele Orman ve XGBoost modellerinin, duygu analizi özelliklerinin entegrasyonu ile birlikte tahmin doğruluğunda belirgin bir artış sağladığı gözlemlenmiştir. Buna karşılık, SVR modelinde kısıtlı düzeyde bir performans artışı saptanmış, Lineer Regresyon modelinde ise duygu skorlarının eklenmesiyle aşırı öğrenme (overfitting) eğilimi gözlemlenmiştir. Bu durum, bazı modellerin duygu verisini etkin biçimde işleyemediğini ya da bu verilerle yüksek genelleme kapasitesine ulaşamadığını göstermektedir.

Elde edilen bulgular, yatırımcı duyarlılığının Bitcoin fiyatları üzerinde etkili olabileceğine ve bu duyarlılığın metin madenciliği teknikleri aracılığıyla nicel hale getirilerek tahmin modellerine entegre edilmesinin öngörü gücünü artırabileceğine işaret etmektedir. Bu bağlamda, çalışmanın kripto para piyasalarında davranışsal finans unsurlarının dikkate alındığı modelleme yaklaşımlarına katkı sağladığı söylenebilir.

Gelecekte yapılacak çalışmaların, BERT gibi gelişmiş doğal dil işleme (Natural Language Processing - NLP) modellerini içeren derin öğrenme temelli yaklaşımlar üzerine yoğunlaşması, farklı coğrafyalardan elde edilen metin veri kümeleriyle modelin genelleme yeteneğinin sınanması ve zaman çözünürlüğü daha yüksek (örneğin saatlik veya dakikalık) veri setleriyle çalışılması önerilmektedir. Ayrıca, bu tür duygu tabanlı tahmin sistemlerinin ekonomik etkilerinin de detaylı olarak analiz edilmesi hem akademik literatür hem de pratik uygulamalar açısından önemli katkılar sağlayacaktır.

Kaynakça

- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40-79. <https://doi.org/10.1214/09-SS054>
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8. <https://doi.org/10.1016/j.jocs.2010.12.007>
- Buathong, W., Sieng-EK, P., & Jarupunphol, P. (2023). Measuring the performance of machine learning forecasting models to support bitcoin investment decisions [J]. *Journal of Data Science and Intelligent Systems*.
- Catalini, C., & Gans, J. S. (2016). Some simple economics of the blockchain. National Bureau of Economic Research (NBER) Working Paper No. 22952. <https://doi.org/10.3386/w22952>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
- Corbet, S., Lucey, B., Urquhart, A., & Yarovaya, L. (2019). Cryptocurrencies as a financial asset: A systematic analysis. *International Review of Financial Analysis*, 62, 182-199. <https://doi.org/10.1016/j.irfa.2018.09.003>
- Dyhrberg, A. H. (2016). Bitcoin, gold and the dollar – A GARCH volatility analysis. *Finance Research Letters*, 16, 85-92. <https://doi.org/10.1016/j.frl.2015.10.008>
- Garcia, D., Tessone, C. J., Mavrodiev, P., & Perony, N. (2014). The digital traces of bubbles: Feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of the Royal Society Interface*, 11(99), 20140623. <https://doi.org/10.1098/rsif.2014.0623>
- Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

- Kaggle. (2025a). Cryptonews 2019–2023 [Veri seti]. Kaggle. Erişim tarihi: 10 Ağustos 2025, <https://www.kaggle.com/datasets/larysa21/cryptonews-2019-2023>
- Kaggle. (2025b). Crypto currencies daily prices [Veri seti]. Kaggle. Erişim tarihi: 10 Ağustos 2025, <https://www.kaggle.com/datasets/svaningelgem/crypto-currencies-daily-prices?select=BTC.csv>
- Kaminski, J. (2014). Nowcasting the Bitcoin market with Twitter signals. *arXiv preprint, arXiv:1406.7577*.
- Kim, Y., Kang, S., & Kim, S. (2016). Financial news analysis using machine learning techniques: An ensemble approach. *Expert Systems with Applications*, 62, 205–215.
- Kim, K. (2020). Cryptocurrency trader's risky decision on social sentiment. In *GITMA 2020: 19th Global Information Technology Management Association (Virtual Conference)*.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137–1145).
- Kristoufek, L. (2013). Bitcoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3(1), 3415. <https://doi.org/10.1038/srep03415>
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1–167. <https://doi.org/10.2200/Soo416ED1Vo1Y201204HLT016>
- Mai, F., Shan, Z., Bai, Q., Wang, X., & Chiang, R. H. L. (2018). How does social media impact Bitcoin value? A test of the silent majority hypothesis. *Journal of Management Information Systems*, 35(1), 19–52. <https://doi.org/10.1080/07421222.2018.1440774>
- Mao, H., Counts, S., & Bollen, J. (2015). *Quantifying the effects of online bullishness on international financial markets* (No. 9). ECB Statistics Paper.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis (5th ed.)*. John Wiley & Sons.
- Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653–7670. <https://doi.org/10.1016/j.eswa.2014.06.009>

- Phillips, R. C., & Gorse, D. (2017). Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *2017 IEEE symposium series on computational intelligence (SSCI)* (pp. 1-7). IEEE. <https://doi.org/10.1109/SSCI.2017.8280809>
- Şahinaslan, Ö., Dalyan, H., & Şahinaslan, E. (2022). Naive bayes sınıflandırıcısı kullanılarak youtube verileri üzerinden çok dilli duygu analizi. *Bilişim Teknolojileri Dergisi*, 15(2), 221-229.
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199-222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- Yermack, D. (2024). Is Bitcoin a real currency? An economic appraisal. In *Handbook of digital currency* (pp. 29-40). Academic Press.