*Avrupa Bilim ve Teknoloji Dergisi*
*Sayı 54, S. 161-173, Aralık 2024*

***Araştırma Makalesi***

www.ejosat.com ISSN:2148-2683

*European Journal of Science and Technology*
*No. 54, pp. 161-173, December 2024*

***Research Article***

# Konu Modelleme Yöntemleri ile Kripto Para Duygu Analizi

Gül Cihan Habek[1*], Mansur Alp Toçoğlu[2], Aytuğ Onan[3]

[1*] Karamanoğlu Mehmetbey Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Karaman, Türkiye, (ORCID: 0000-0003-1748-3486), gulhabek@kmu.edu.tr

[2] Katip ÇelebiUniversity, Faculty of Engineering and Architecture, Departmant of Computer Engineering, İzmir, Turkey, (ORCID: 0000-0003-1784-9003), mansur.alp.tocoglu@ikcu.edu.tr

[3] Katip ÇelebiUniversity, Faculty of Engineering and Architecture, Departmant of Computer Engineering, İzmir, Turkey, (ORCID: 0000-0002-9434-5880), aytug.onan@ikc.edu.tr

(İlk Geliş Tarihi 29 Ocak 2024 ve Kabul Tarihi 16 Ekim 2024)

(**DOI:** 10.5281/zenodo.14578628)

**ATIF/REFERENCE:** Habek, G. C., Toçoğlu, M. A. & Onan, A. (2024). Konu Modelleme Yöntemleri ile Kripto Para Duygu Analizi. *Avrupa Bilim ve Teknoloji Dergisi*, (54), 161-173.

**Öz**

İnternet ortamında kullanılmak üzere tasarlanmış olan kripto paralar çıktığı günden bu yana pek çok insanın ilgisini çekmeyi başarmıştır. Kripto para ticaretine yönelen insan sayısı arttıkça konu ile ilgili metin sayısı da artmış ve büyük veri setine dönüşen metinlerin analiz edilme ihtiyacı doğmuştur. Literatürde farklı konular ile ilgili Twitter platformundan çekilen tweetler kullanılarak yapılan analiz çalışmaları mevcuttur. Bu çalışma kripto paraların yükselen popülerliği ve internet kullanıcılarının bu alana olan ilgisi doğrultusunda, Twitter platformunda paylaşılan kripto para etiketli tweetler üzerinden duygu analizi gerçekleştirmeyi amaçlamaktadır. Twitter, geniş veri seti sunan popüler bir sosyal medya platformu olduğundan kullanıcıların kripto paralarla ilgili duygusal ifadelerini içeren tweetler, bu çalışmanın temel veri kaynağını teşkil etmekte olup tweetler LSTM ile otomatik etiketlenerek çalışmaya özgü bir veri seti oluşturulmuştur. Konu modelleme yöntemlerinden LDA ve NMF algoritmaları ile özellik tespiti yapılmış; NB, LR ve DVM algoritmaları ile modeller oluşturulup sınıflandırma doğruluğu ve f1-ölçütü metrikleri için sonuçlar alınmıştır. Alınan sonuçlar karşılaştırıldığında LDA yöntemi ve DVM sınıflandırıcısı kullanıldığında en yüksek başarı elde edilmiş olup, konu özellik boyutu 500 ve bileşen sayısı 40 olarak ayarlanarak oluşturulan modelde %87.89 doğruluk ve %87.85 F1-ölçütü değerlerine ulaşılmıştır. NMF yöntemi ile elde edilen en iyi sonuç da konu özellik boyutu 500 ve bileşen sayısı 50 olduğunda DVM sınıflandırıcısı ile oluşturulan model ile %87.43 doğruluk ve %87.34 F1-ölçütü olarak alınmıştır. İkinci en iyi performans ise her iki yöntem için de LR sınıflandırıcısı ile elde edilmiştir; LDA yöntemi ile konu özellik boyutu ve bileşen sayısı sırası ile 300 ve 40 seçildiğinde %87.13 doğruluk, 500 ve 30 seçildiğinde %87.13 F1-ölçütü değerleri kaydedilirken NMF yöntemi ile konu özellik boyutu 500 ve bileşen sayısı 50 olduğunda %87.01 doğruluk ve %86.90 F1-ölçütü değerlerine ulaşılmıştır. En düşük performans ise her iki yöntem için de NB sınıflandırıcısı ile alınmış, NB sınıflandırıcı en iyi performansını konu özellik boyutu ve bleşen sayısı sırasıyla 200 ve 50 seçildiğinde LDA ile %73.00 doğruluk ve %73.01 F1-ölçütü, 500 ve 40 seçildiğinde NMF ile %70.00 doğruluk ve %70.09 F1-ölçütü değerleri ile göstermiştir. Sonuçlara göre hem LDA hem de NMF yöntemleri genel olarak başarılı performans göstermiştir. Her iki yöntemde de en yüksek ve en tutarlı performanslar DVM sınıflandırıcısı ile alınmış olup, DVM algoritmasının kullanılan diğer iki sınıflandırıcıya kıyasla, en başarılı sınıflandırıcı olduğu sonucuna varılmıştır. Ayrıca çalışmadan elde edilen bulgular, Twitter'daki kripto paralarla ilgili duygusal ifadelerin etkili bir şekilde analiz edilebileceğini göstermektedir.

**Anahtar Kelimeler:** Kripto Para, Duygu Analizi, LSTM ile Otomatik Etiketleme, Konu Modelleme Yöntemleri, Metin Sınıflandırma.

# Cryptocurrency Sentiment Analysis with Topic Modeling Methods

**Abstract**

Cryptocurrencies designed to be used on the internet have attracted the attention of many people since the day they were released. As the number of people turning to cryptocurrency trading increased, the number of texts on the subject also increased and the need to analyze the texts, which turned into a large data set, arose. There are analysis studies in the literature using tweets taken from the Twitter platform on different topics. This study aims to perform sentiment analysis on cryptocurrency-tagged tweets shared on the Twitter

[*] Sorumlu Yazar: gulhabek@kmu.edu.tr

platform, in line with the rising popularity of cryptocurrencies and the interest of internet users in this field. Since Twitter is a popular social media platform that offers a wide data set, tweets containing users' emotional expressions about cryptocurrencies constitute the main data source of this study, and a data set specific to the study was created by automatically tagging tweets with LSTM. Feature detection was made with LDA and NMF algorithms, which are topic modeling methods; Models were created with NB, LR and SVM algorithms and results were obtained for classification accuracy and f1-score metrics. When the results were compared, the highest success was achieved when the LDA method and SVM classifier were used, and 87.89% accuracy and 87.85% F1-criterion values were reached in the model created by setting the subject feature size as 500 and the number of components as 40. The best result obtained with the NMF method was 87.43% accuracy and 87.34% F1-criterion with the model created with the SVM classifier when the subject feature size and the number of components were 500 and 50, respectively. The second-best performance was obtained with the LR classifier for both methods; With the LDA method, 87.13% accuracy was recorded when the subject feature size and the number of components were selected as 300 and 40, and 87.13% F1-criterion values were recorded when 500 and 30 were selected, respectively, while with the NMF method, 87.01% accuracy and 86.90% F1-criterion values were reached when the subject feature size was 500 and the number of components was 50. The lowest performance was obtained with the NB classifier for both methods. The NB classifier showed its best performance with 73.00% accuracy and 73.01% F1-measure values with LDA when the subject feature size and the number of components were selected as 200 and 50, respectively, and 70.00% accuracy and 70.09% F1-measure values with NMF when 500 and 40 were selected. According to the results, both LDA and NMF methods generally showed successful performance. The highest and most consistent performances were obtained with the SVM classifier in both methods, and it was concluded that the SVM algorithm was the most successful classifier compared to the other two classifiers used. In addition, the findings obtained from the study show that emotional expressions related to cryptocurrencies on Twitter can be analyzed effectively.

**Keywords:** Cryptocurrency, Sentiment Analysis, Automatic Tagging with LSTM, Topic Modeling Methods, Text Classification.

# 1. Giriş

2000'li yıllarda sınırlı bir kullanım alanına sahip olan internet, gün geçtikçe popülerliğini artırmaktadır. Teknolojik ilerlemelerle birlikte, internet üzerinden gerçekleştirilen işlemlerin sayısı artmakta ve her alanda kullanılmaya başlanmaktadır. İlk zamanlarda daha çok bilgiye erişim ve temel iletişim ihtiyaçlarını karşılamak için kullanılan internet, yaygınlaşması ve gelişimi sayesinde günümüzde alışveriş, eğitim-öğretim ve bankacılık gibi birçok işlemin dijital ortamlarda kolay ve hızlı bir şekilde gerçekleştirilmesine olanak tanımaktadır (Böttger, 2017).

Blockchain teknolojisi üzerine kurulu olan kripto paralar, internet ortamında kullanılmak üzere tasarlanmış ve çıktığı günden itibaren devletler, kurumlar ve bireylerin dikkatini çekmeyi başarmış merkeziyetsiz dijital varlıklardır (Habek et al., 2022). Geleneksel finans sistemlerinden farklı olarak, herhangi bir merkezi otoriteye bağlı olmayan bu dijital varlıklar, kullanıcılar arasındaki işlemlerin aracısız ve şeffaf bir şekilde gerçekleştirilmesine olanak tanır (Shoetan & Familoni, 2024). Kripto paraların temelini oluşturan Blockchain teknolojisi, bilginin hızlı ve güvenilir bir şekilde iletilmesini sağlayan, her biri bir önceki ile bağlantılı olan bloklardan oluşan bir zincir yapısını içerir. Kripto paraların güvenliğini sağlamak için kullanılan kriptoloji tekniği ise veriyi okunamaz hale getirerek, alıcısı dışında herhangi birinin ele geçirmesi durumunda anlaşılamaz hale gelmesini sağlayan bir matematiksel hesaplama sürecini kapsar. Bu şifreleme tekniği hem verilerin gizliliğini hem de bütünlüğünü koruma altına alır. Kripto paraların benzersiz özellikleri, bu teknolojileri kullanmalarından kaynaklanmaktadır. Hızlı ve güvenilir işlem yapma yetenekleri, Blockchain ve kriptoloji tekniklerinin etkili bir şekilde kullanılmasıyla mümkün olmaktadır. Kripto para birimleri ile gerçekleştirilen işlemler, geleneksel finans sistemlerine göre daha güvenli ve takip edilemez bir yapıya sahiptir.

Blockchain'in merkeziyetsiz yapısı sayesinde, işlemler küresel çapta herhangi bir coğrafi sınır olmaksızın gerçekleştirilebilirken, kriptografi sayesinde de bu işlemler güvence altına alınır (Shoetan & Familoni, 2024). Bu sayede, kripto paralar finans sektöründe önemli değişikliklere yol açmanın yanı sıra; teknoloji, sağlık, enerji ve daha birçok sektörde de dikkate değer yeniliklere imza atmıştır. Her geçen gün daha fazla benimsenen kripto paralar, geleceğin dijital ekonomisinin temel yapı taşlarından biri olarak kabul edilmektedir.

Son yıllarda kripto para birimlerine gösterilen ilginin artması ile çok fazla metinde ismi geçmeye başlamıştır. Konu ile ilgili metin sayısının artması kripto para ticareti ile ilgilenen bireylerin paylaşılan metinleri analiz etme isteğini doğurmuştur. Analiz için Twitter üzerinden paylaşılan konu ile ilgili ham veriler piyasa ile ilgilenen kişi ve kurumların duygu ve düşüncelerini ölçmek için önemli kaynaklardır. Aktif kullanıcı ve paylaşımların fazla olması platformu büyük bir veri kaynağına sahip hale getirmektedir. Bu durum araştırmacıların belirli bir konudaki Twitter paylaşımlarının duygularını analiz etmek istemesinin en büyük nedeni olarak gösterilebilir.

Kripto paralarla ilgili veri analiz süreçleri, genellikle sosyal medya duygu analizi yöntemleri kullanılarak gerçekleştirilmektedir. Bu yöntemler kullanıcıların tweet'lerinde yer alan olumlu, olumsuz veya nötr ifadeleri belirleyerek piyasa beklentilerini tahmin etmeye yardımcı olur. Özellikle Twitter'daki büyük hacimli verinin işlenmesi, doğal dil işleme teknikleri ve makine öğrenimi algoritmalarıyla mümkün hale gelmiştir. Kripto para piyasasındaki fiyat dalgalanmaları genellikle kullanıcıların bu platformlardaki duygu durumlarına bağlı olduğundan sosyal medyada duygu analizi yatırımcılara ve piyasa gözlemcilerine stratejik avantajlar sunar.

Bu bağlamda kripto para birimleri ile ilgili Twitter paylaşımlarını analiz etmek, piyasa hareketlerini önceden tahmin etmek ve risk yönetimi stratejileri geliştirmek isteyenler için önemli bir araç haline gelmiştir. Analizlerin sonuçları yatırımcıların bilinçli kararlar almalarına yardımcı olurken kripto para piyasalarındaki ani dalgalanmaları da daha anlaşılır kılmaktadır.

Duygu analizi ve sınıflandırma çalışmalarında sıklıkla tercih edilen konu modelleme yöntemi, ele alınan veri setindeki dokümanların konularını belirlemek için kullanılan, dokümanlar üzerinde analiz yapan denetimsiz bir modeldir (Abdelrazek et al.,

2023). Bu sayede, büyük hacimli veri setleri anlamlandırılabilir hale gelir, ayrıca metinlerin sınıflandırılması ve ilgili duyguların tespiti kolaylaşır.

Literatürde geniş yer bulan ve yaygın olarak kullanılan birçok konu modelleme tekniği bulunmaktadır. En bilinen ve sıklıkla uygulanan yöntemler Gizli Dirichlet Tahsisi (Latent Dirichlet Allocation, LDA), Gizli Semantik Analiz (Latent Semantic Analysis, LSA), Negatif Olmayan Matris Faktoringi (Non-Negative Matrix Factorization-NMF), Olasılıksal Gizli Semantik Analiz (Probabilistic Latent Semantic Analysis, PLSA), İlişkili Konu Modeli (Correlated Topic Model-CTM), Hiyerarşik Dirichlet Süreci (Hierarchical Dirichlet Process-HDP) ve Yapısal Konu Modeli (Structural Topic Model-STM) yaklaşımlarıdır. LDA, her bir dokümanın çeşitli konuların karışımından oluştuğunu ve her konunun da belirli kelimelerden meydana geldiğini varsayarak, belgelerdeki gizli konuları tespit etmeye çalışır. LSA, kelimeler arasındaki anlamsal benzerlikleri ortaya koyarak metinler arasındaki ilişkileri ve konusal yapıları analiz eder. NMF, metinlerin kelime-doküman matrislerini pozitif faktörlere ayırarak metinlerdeki gizli konuları ortaya çıkarır. NMF, LDA ve LSA gibi tekniklerle kıyaslandığında daha hızlı ve anlaşılır sonuçlar vermektedir. Bu temel yöntemlerin yanı sıra, literatürde yer alan diğer gelişmiş yaklaşımlar olan PLSA, CTM, HDP ve STM ise, özellikle çoklu değişkenlerin incelendiği, ilişkili konuların bulunduğu veya veri setinin karmaşıklığının arttığı durumlarda kullanılır.

Konu modelleme yöntemi üzerine yapılan çalışmalara bakıldığında, bu yöntemlerin çeşitli alanlarda başarıyla uygulandığı görülmektedir. İlgili çalışmalara bakıldığında, çalışmalardan birinde konu tespit sistemi geliştirmek amacı ile Siyaset, Ekonomi, Spor ve Kültür haberleri olmak üzere toplam dört farklı kategori için toplam 400.000 veri toplanarak Gizli Dirichlet Ayırımı (GDA) yöntemi ile model eğitilmiştir (Aydın & Hallaç, 2021). Eğitilen model veri setinde bulunmayan bir dokümanın konu tespitini %94.2 yüksek doğruluk oranı ile yapabilmektedir.

Sonraki çalışmada www.otelpuan.com sitesi üzerinden web crawler kullanılarak 1000 otel müşterisinin yorumları toplanılmış ve toplamda 5364 cümle ile duygu analizi çalışması yapılmıştır (Ekinci & Omurca, 2017). Bir önceki çalışmada olduğu gibi Gizli Dirichlet Ayırımı (GDA) yöntemi tercih edilmiş ve oluşturulan modelin performansı kesinlik, duyarlılık ve F-Ölçütü olmak üzere 3 farklı metrik ile ölçülmüştür. En başarılı sonuç duyarlılık metriği ile %80 olarak elde edilmiştir.

Bir diğer çalışmada sosyal haber ve tartışma sitesi olan Reddit üzerinden gönderilen kanser hastalığı ile ilgili gönderiler çekilmiş ve yine Gizli Dirichlet Ayırımı (GDA) yöntemi ile en çok paylaşım yapılan konu başlıkları elde edilmiştir (Altıntaş et al., 2021). Çalışmanın devamında içerik analizi yapmak adına konu başlıklarının konuya olan yakınlığı ele alınmıştır. Çalışmanın sonucunda oluşturulan model ile konu başlıklarında kullanılan kelimelerin konu içeriği ile uyumlu olduğu sonucuna varılmıştır.

Konu modelleme tekniğini kullanarak sınıflandırma yapılması amaçlanan bir çalışmada ise şikayetvar.com sitesi üzerinden gönderilen Teknosa müşterilerinin şikayetleri analiz edilmiştir (Koruyan, 2022). Çalışmada konu modelleme teknikleri arasından BERTTopic kullanılarak oluşturulan model ile alınan sonuçlara göre en çok şikâyet alan konular bilgisayar, telefon, televizyon, tablet, kulaklık, kargo, sipariş iptali ve mağaza çalışanlarının tutumu olmuştur.

RapidMiner ile Twitter verilerinin konu modellemesi adlı çalışmada geleneksel yöntemler ile analiz edilmesi oldukça zor olan büyük veriler için yeni bir yöntem geliştirilerek verilerin daha kolay bir şekilde analiz edilebilmesi amaçlanmıştır (Ankaralı & Külcü, 2020). Bunun için popüler sosyal medya platformlarından biri olan Twitter üzerinden gönderilen "ordu" kelimesi içeren tweetler analiz edilmiştir. Çalışmanın ikinci aşamasında içerisinde "Hacettepe" kelimesi geçen tweetlerin konu modellemesi Latent Dirichlet Allocation (LDA) algoritması kullanılarak yapılmıştır.

Kurumların üniversite bilgi yönetim sistemi, kısaca ÜBYS için talep ettikleri servis destek talepleri için konu modellemesinin yapıldığı çalışmada ÜBYS içerisinde bulunan "öğrenci bilgi sistemi-AIS", "personel bilgi sistemi-HRM", "elektronik belge yönetim sistemi-ERMS" ve "bilimsel araştırma projeleri-SRP" sistemleri için talep edilen servis destek taleplerinin konu modellemesi Latent Dirichlet Allocation (LDA) kullanılarak yapılmıştır (Onan et al., 2020). Metin ön işleme adımlarından geçirilen 4 konulu ve 8 konulu servis talepleri için oluşturulan modeller kullanıldığında uyumlu bir veri seti elde edilmiştir.

Twitter gönderileri üzerinden duygu analizi yapılmak istenen bir çalışmada (Onan & Bayar, 2017) Twitter API kullanılarak 5300 olumlu ve 5300 olumsuz veri çekilmiş Latent Dirichlet Allocation (LDA) yöntemi kullanılarak veriler temsil edilmiştir. Farklı geleneksel algoritmalar kullanılarak sonuçlar karşılaştırıldığında 50 konu seçildiğinde Naive Bayes algoritması ile en yüksek sonuç alınmıştır.

Son çalışmada ise Twitter platformu üzerinden çevreci sivil toplum kuruluşlarının 2020 ile 2021 arasında yapmış oldukları iklim ile ilgili 7764 paylaşım çekilmiş; paylaşımların konu dağılım oranlarını hesaplamak amacıyla metinsel içerik analizi Latent Dirichlet Allocation (LDA) yöntemi ile gerçekleştirilmiştir (Günay & Güçdemir, 2020).

Bu çalışmada konu modelleme yöntemleri kullanılarak Kripto paralar ile ilgili bir duygu analizi çalışması yapılması amaçlanmıştır. Çalışmanın temel amacı, Twitter üzerinden toplanan etiketli veriler kullanarak bir duygu analizi modeli geliştirmek ve bu model aracılığıyla kripto para piyasasıyla ilgili paylaşımları sınıflandırarak duygu analizini gerçekleştirmektir. Çalışmanın veri toplama aşamasında, Twitter platformunda kripto paralarla ilgili belirli anahtar kelimeler ve özel etiketler kullanılarak veri toplanmıştır. Toplanan verilerin otomatik olarak etiketlenmesi için gelişmiş Derin Öğrenme yöntemlerinden biri olan Uzun Kısa Süreli Bellek (Long Short-Term Memory, LSTM) ağı kullanılmıştır. LSTM, zaman serisi verilerinin ve ardışık bilgilerin işlenmesinde etkili bir model olup bu çalışmada etiketleme işleminin otomasyonu için kullanılmıştır (Huang et al., 2015). Böylece çalışmaya özgü bir veri seti oluşturulmuş ve konu modelleme ve duygu analizi aşamalarında bu veri seti kullanılmıştır. Veri setinin oluşturulması ve içeriği ile ilgili detaylı bilgiye Bölüm 2'de yer verilmiştir. Çalışmada ayrıca Term Frequency (TF) ağırlıklandırma ve n-gram kelime temsil yöntemi kullanılmıştır. Özelliklerin tespiti için literatürde sıklıkla tercih edilen LDA ve NMF konu modelleme yöntemleri kullanılmış olup geleneksel sınıflandırma algoritmalarından Naive Bayes (NB), Lojistik Regresyon (LR) ve Destek Vektör Makinesi (DVM) kullanılarak

sınıflandırma doğruluğu ve f1-ölçütü metrikleri için sonuçlar alınmıştır. Kullanılan algoritmalar ve metrikler ile ilgili detaylı açıklamalar Bölüm 2'de yapılmıştır. Alınan sonuçlara göre, genel olarak iyi performans değerleri elde edilmiştir. En yüksek başarı, konu özellik boyutu 500 ve bileşen sayısı 40 olarak ayarlandığında, LDA yöntemi ve DVM sınıflandırıcısı kullanılarak oluşturulan modelde %87.89 doğruluk ve %87.85 F1-ölçütü ile elde edilmiştir. Ayrıca NMF yöntemi ile de en yüksek performans değerleri, DVM sınıflandırıcısı kullanılarak sağlanmış; konu özellik boyutu 500 ve bileşen sayısı 50 parametreleri için %87.43 doğruluk ve %87.34 F1-ölçütü sonuçları elde edilmiştir.

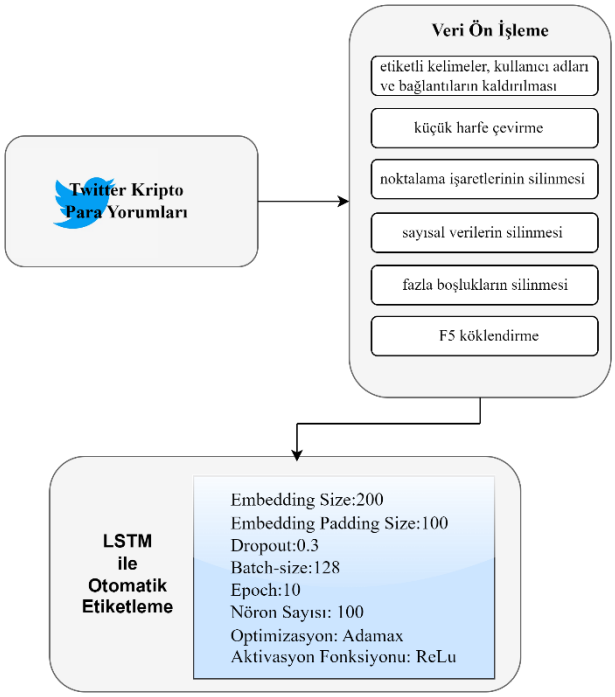
# 2. Materyal ve Metot

Bu bölümde çalışmada kullanılan veri setinin oluşturulması süreci, metin temsil yöntemleri, özellik tespitinde kullanılan konu modelleme yöntemleri, sınıflandırma aşamasında kullanılan geleneksel makine öğrenmesi algoritmaları ve performansları değerlendirirken kullanılan performans metriklerinin teorik açıklamalarına yer verilmiştir.

## 2.1. Veri Setinin Oluşturulması

Twitter platformu üzerinden paylaşılan 21.03.2022-14.11.2022 tarihleri arasındaki #kripto, #bitcoin, #ethereum, #eth, #defi, #nft, #btc, #bitci, #avalanche, #avax, # chiliz, #chz, #xrp, #solana etiketli Türkçe dilindeki 218.226 tweet, Twitter'dan metin tabanlı veri çekmek için API sınırlamaları olmadan geniş ve kapsamlı veri toplama imkânı sunan etkili bir araç olan Snscrape kazıyıcı kullanılarak çekilmiştir. Ön işleme aşamasında Şekil 1'de görüldüğü üzere verilerdeki etiketli kelimeler, kullanıcı adları ve bağlantılar kaldırılmış; tüm harfler küçük harfe çevrilmiş, noktalama işaretleri, sayısal veriler ve fazla boşluklar silinmiştir. Son olarak kök bulma aşamasında F5 köklendirme yöntemi kullanılarak her terim köklerine ayrılmıştır.

Ön işleme aşamasından geçirilen veriler Uzun Kısa Süreli Bellek (Long Short Term Memory, LSTM) derin öğrenme modeli ile otomatik bir şekilde etiketlenmiştir. Otomatik etiketleme, bir veri kümesindeki verilerin belli kategorilere ya da sınıflara ayrılmasını sağlayan bir süreçtir. Manuel etiketleme genellikle zaman alıcı ve hata yapmaya açık olduğu için çalışamada LSTM ile otomatik etiketleme işlemi gerçekleştirilmiş olup bu sayede etiketleme işlem hızlandırılmış ve daha tutarlı hale getirilmiştir.




*Şekil 1. Veri Setinin Temizlenmesi ve Etiketlenmesi Adımları*

Otomatik etiketleme için oluşturulan LSTM modelinde Şekil 1'de görüldüğü gibi Gömme Boyutu (Embedding Size) 200, Gömme Dolgu Boyutu (Embedding Padding Size) 100, Bırakma (Dropout) 0.3, Batch Boyutu 128 ve devir sayısı (epoch) 10 olarak girilmiştir. Modelin karmaşık ilişkileri öğrenebilmesi için LSTM'in nöron sayısı 100 olarak seçilmiş olup "Adam" optimizasyon algoritmasının bir varyasyonu olan ve yüksek boyutlu veri setlerinde güzel sonuçlar veren "Adamax" optimize edici ve "ReLu (Rectified Linear Unit)" aktivasyon fonksiyonu kullanılmıştır. Oluşturulan sınıflandırma modelinin eğitiminde manuel etiketli veri seti [1,2] kullanılmış olup 218.226 etiketli veri arasından rastgele 99.980 veri seçilerek bu çalışmada kullanılacak veri seti oluşturulmuştur. Veri setindeki sınıfların dağılımı Tablo 1'de gösterilmiştir. Otomatik etiketli veri seti ile ilgili örnek verilere ise Tablo 2'de yer verilmiştir.

*Tablo 1. Sınıflardaki Tweet Sayıları*

| Sınıflar | Tweet Sayıları |
|---|---|
| Pozitif | 49.993 |
| Negatif | 49.987 |
| **Toplam** | 99.980 |

*Tablo 2. Otomatik Etiketli Veri Örnekleri*

| ID | Metin (Tweet) | Sınıf |
|---|---|---|
| 1 | xrp bu hafta guzel hareketli beklentiler yüksek sende kazanca ortak ol | 1 |
| 2 | bitcoin ani hareketleri altları eziyor bitcoin ay kapanışı için kendini yükseklere atmaya çalışıyor yarın ay kapanışı btc de geri çekilme gelebilir temkinli olmakta fayda var nakitte kalmanızı tavsiye ederim | 0 |
| 3 | kripto para fake artık bunun borsa ile bir alakası filan yok ayı dönemi bitmeyecek | 0 |
| 4 | piyasaya para akıyor farkındasınız değil mi dominance düşüyor total yükseliyor btc yükseliyor | 1 |
| 5 | ben pumplanmışken ancak yakalayabildim fiyat düşüşünü görünce de trene atladım atlamamak mümkün değil zira beraber uçacağız | 1 |

## 2.2. Metin Temsil Yöntemleri

Metin temsil yöntemleri, doğal dil işleme alanında metin verilerini bilgisayarlar tarafından işlenebilmesi ve analiz edilebilmesi amacıyla sayısal formata dönüştüren yöntemlerdir (Das & Chakraborty, 2018). Bu yöntemler, metin verisinin içeriğini matematiksel ve istatistiksel modellere uygun hale getirmek için kullanılır. Bu çalışmada, metin verilerinin temsil edilmesi aşamasında doğal dil işleme alanında sıklıkla kullanılan n-gram kelime temsil yöntemi tercih edilmiştir. N-gram yönteminde sıralı harfler, heceler veya kelimeler n sayısına bağlı olarak gruplara ayrılmaktadır. "N" sayısı, kaç öğenin bir araya gelerek bir grup oluşturacağını belirler. 1-gram (unigram) yalnızca tek bir kelimeyi, 2-gram (bigram) ardışık iki kelimeyi, 3-gram (trigram) ise ardışık üç kelimeyi içeren bir grup oluşturur. Bu çalışmada kelime bazlı unigram temsil yönteminden yararlanılmıştır. Veri setinde bulunan bir doküman üzerinden örnek vermek gerekirse “dijital para çağı başlıyor” cümlesinde kelimeler kelime bazlı unigram temsil yöntemi ile “dijital”, “para”, “çağı”, “başlıyor” şeklinde gruplara ayrılmıştır.

Kelimelerin ağırlıklandırması için ise doğal dil işleme ve metin sınıflandırma çalışmalarında performansı arttırmak amacıyla kullanılan geleneksel yöntemlerden Terim Frekansı (Term Frequency, TF) yöntemi tercih edilmiştir (Çoban & Tümüklü Özyer, 2018). TF yönteminde bir dokümanda sıklıkla geçen kelimeler daha büyük öneme sahiptir ve bir terimin bir dokümanda bulunma sayısı ile doğru orantılı olarak ağırlığı atanır.

Sonuç olarak, Şekil 2’de verilen mimaride görüldüğü üzere bu çalışmada n-gram ve TF yöntemleri bir arada kullanılarak kelimelerin temsil edilmesi ve bu temsillerin ağırlıklandırılması sağlanmıştır. Bu yöntemler, sınıflandırma modellerinin eğitimi sırasında metinlerin daha verimli bir şekilde analiz edilmesine olanak tanımış ve modelin performansını artırmıştır.

## 2.3. Konu Modelleme Algoritmaları

Konu modelleme yöntemleri metin verilerinin anlamsal yapısını belirleyen denetimsiz makine öğrenmesi teknikleridir. (Ekinci & Omurca, 2017). Bu yöntemler, özellikle büyük ölçekli metin veri setlerinde gizli kalıpları ortaya çıkarmak için kullanılmaktadır (Turan et al., 2023). Çalışmada özelliklerin çıkarılması adımında Şekil 2’de görüldüğü gibi konu modelleme yöntemlerinden olan LDA ve NMF algoritmalarından yararlanılmış olup bu kısımda algoritmalar ile ilgili teorik bilgilere yer verilmiştir.

### *2.3.1. Gizli Dirichlet Tahsisi (Latent Dirichlet Allocation-LDA) Algoritması*

LDA belirlenen konu sayısı ve sınıflara göre konulara etiket ataması yapan istatistiksel ve denetimsiz makine öğrenmesi yöntemidir (Ekinci & Omurca, 2017). Metin verilerinde bulunan anlamsal gizli terimleri ortaya çıkarmak amacıyla 2003 yılında Blei ve ark. tarafından geliştirilmiştir (Blei et al., 2003). LDA algoritması öncelikle veri setinde bulunan dokümanların kelimelerine rastgele konular atayarak başlar. Olasılıksal bilgi için lokal olasılık ve global olasılık bilgileri toplanılır. Lokal olasılıkta her doküman için konulara atanan kelime sayıları tutulurken global olasılıkta dokümanların genelinde konulara atanan kelime sayıları tutulmaktadır. Elde edilen olasılıksal bilgilerden sonra dokümanlardaki kelimelere tekrar konu atamaları yapılır ve bu işlemler girilen iterasyon sayısı kadar tekrarlanır.

LDA'nın güçlü yanı, metin verilerindeki gizli konuları ve bu konuların kelimelerle ilişkisini ortaya çıkarmasıdır. Bu avantajı sayesinde metin madenciliği, bilgi çıkarımı, sınıflandırma problemleri gibi birçok farklı uygulamada tercih edilmektedir. Olasılıksal temelli olması sayesinde metinler arasındaki belirsizlik ve varyasyon dikkate alınır. Algoritma, her dokümanın bir dizi konu ile temsil edilmesini ve her kelimenin bir konuya ait olasılık dağılımını ilişkilendirir.

### *2.3.2. Negatif Olmayan Matris Faktoringi (Non-Negative Matrix Factorization-NMF) Algoritması*

Bir diğer istatistik tabanlı popüler konu modelleme algoritması olan NMF, girdi gövdesinin boyutunu küçülten ve veri analizinde kullanılan bir modeldir (M'sik & Casablanca, 2020). Bu algoritma Denklem 1'de verildiği gibi belge terim matrisini daha düşük boyutlu iki matrise ayrıştırarak veriyi daha yönetilebilir bir forma sokar (Wang, 2023).

$$V_{mxn} = W_{mxk} * H_{kxn} \qquad (1)$$

Denklem 1'de görüldüğü gibi NMF algoritması belge terim matrisi V'yi, satırları kelimelerden oluşan H matrisi ve sütunları kelimelerin cümlelerdeki ağırlıklarını tutan W matrisine ayrıştırmıştır. Denklem 1'de matrislerin boyutlarını temsil eden m kelime sayısını, k konu sayısını ve n toplam doküman sayısını tutmaktadır.

NMF, verilerin altında yatan yapıları keşfetmek ve bu yapıların yorumlanabilir temsillerini oluşturmak için kullanılır. Özellikle metin madenciliğinde, belgelerdeki gizli konuları bulmak için etkili bir tekniktir (Lee & Seung, 1999).

NMF'nin avantajı verilerin ayrıştırılmasında negatif değerlere izin vermemesidir. Bu sayede elde edilen faktörlerin yorumlanması kolaylaşır ve her faktör ya da konu sadece pozitif değerlere sahip bileşenlerden oluşur. Bu durum özellikle metin analizlerinde her konunun belirli kelimelerle pozitif olarak ilişkilendirilmesini garantiler.

## 2.4. Makine Öğrenmesi Algoritmaları

Günümüzde birçok disiplinde kullanılan makine öğrenmesi, makinelerin yani bilgisayarların insanların düşünme yeteneğini taklit ederek insanlar gibi karar alma becerisini kazanması demektir. Makine Öğrenmesi algoritmaları denetimli, denetimsiz, yarı denetimli ve pekiştirmeli öğrenme olmak üzere temelde 4 alt başlığa ayrılmaktadır. Denetimli öğrenme modellerinde, eğitim için giriş verilerine karşılık gelen etiketli veriler kullanılmakta ve model bu etiketli veriler üzerinden öğrenim gerçekleştirmektedir.

Literatürde sınıflandırma, regresyon ve kümeleme problemlerinde makine öğrenmesi algoritmalarından sıklıkla yararlanılmaktadır. Sınıflandırma, veri noktalarının önceden belirlenen sınıflardan birine atanması işlemidir. Bu çalışmada ise denetimli makine öğrenmesi algoritmaları kullanılarak Twitter'dan toplanılan verilerin sınıflandırılması yapılmak istenmiştir. Çalışmada Şekil 2'de verilen mimaride görüldüğü gibi geleneksel makine öğrenmesi algoritmalarından Naive Bayes (NB), Lojistik Regresyon (LR) ve Destek Vektör Makinesi (DVM) algoritmaları kullanılarak sınıflandırma doğruluğu (accuracy) ve f1-ölçütü metrikleri ile sonuçlar alınmış olup alınan sonuçlar performans açısından karşılaştırılmıştır. Kullanılan makine öğrenmesi algoritmaları ile ilgili teorik bilgilere bu kısımda yer verilmiştir.

### *2.4.1. Naive Bayes (NB) Algoritması*

Naive Bayes (NB) sınıflandırma algoritması belirlenen sınıflar için girdilerin dağılımını modelleyen denetimli bir makine öğrenmesi algoritmasıdır. Metin sınıflandırması, spam filtreleme ve duygu analizi gibi doğal dil işleme problemlerinde yaygın olarak kullanılmaktadır (Zhang, 2004). Metinlerin sınıflandırılması probleminde ele alınan dokümandaki terimlerin dağılımlarından yola çıkarak bir sınıfın sonuca dayalı olasılığını hesaplar (Dewi et al., 2023). Denklem 2'de verilen Bayes teoremine dayalı olasılıksal bir algoritmadır. Sınıflandırma problemlerini çözmek için verilerin özelliklerine dayalı olarak olasılık hesaplamaları yapar.

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \qquad (2)$$

Denklem 2'de A etiketleri ve B özellikleri temsil eder. Buna göre;

- $P(A|B)$, ele alınan özellik setinin belirli bir etiket ile ilişkili olma olasılığını hesaplamaktadır. Diğer bir ifade ile bir dokümanın belirli bir sınıfa ait olma olasılığını gösterir.
- $P(B|A)$, sınıf (A) gözlemlendiğinde, özellikler kümesi (B)'nin ortaya çıkma olasılığıdır. Kısaca bir sınıfın içinde belirli terimlerin görülme olasılığıdır.
- $P(A)$, sınıf (A)'nın genel olasılığıdır, yani sınıfın veri kümesindeki frekansıdır.
- $P(B)$, özellikler kümesi (B)'nin genel olasılığıdır, yani yani tüm dokümanlarda belirli bir terim kombinasyonunun görülme olasılığıdır.

NB algoritması Denklem 2'den anlaşılacağı üzere dokümanlardaki terimlerin yani kelimelerin dağılımını kullanarak dokümanın belirli bir sınıfa ait olma olasılığını hesaplamaktadır (McCallum & Nigam, 1998). Algoritmanın "Naive (saf)" olarak adlandırılmasının sebebi ise dokümanlar içerisindeki terimlerin bağımsız olarak kabul edilmesidir. Terimlerin bağımsız olması varsayımı, gerçek hayatta genellikle tam anlamıyla doğru olmasa da NB algoritması pratikte oldukça başarılı sonuçlar vermektedir.

NB algoritmasının en büyük avantajlarından biri, büyük veri kümeleriyle ve çok sayıda özelliğin bulunduğu veri setleriyle etkili bir şekilde çalışabilmesidir. Ayrıca, algoritma hem basit hem de hızlıdır, bu nedenle metin sınıflandırması ve diğer doğal dil işleme problemlerinde yaygın olarak tercih edilmektedir.

### *2.4.2. Lojistik Regresyon (LR) Algoritması*

Lojistik regresyon algoritması, bir lojistik (sigmoid) fonksiyon yardımı ile istatistiksel tahminlerde bulunup değişkenlerdeki bağların tespit edilip test edilmesinde kullanılmaktadır (Balli & Sağbas, 2017). Algoritma, bağımlı değişkenin belirli bir sınıfa ait olup olmadığını olasılıksal olarak modellemek için giriş verilerindeki değişkenler arasındaki ilişkileri inceler. Özellikle bağımlı değişkenin iki kategoriye ayrıldığı durumlarda sıklıkla tercih edilmektedir. Değişkenler arasındaki bağımlılıkları analiz etmenin yanı sıra olasılık tahminleri yaparak sınıflandırma işlemi gerçekleştirmektedir.

Uygulanabilirliği ve yorumlanması kolay olduğu için ikili sınıflandırma problemlerinde çokça tercih edilen bir makine öğrenmesi algoritmasıdır (Cox, 1958). Model, lojistik fonksiyon kullanarak çıktı değerlerini [0,1] aralığına çeker ve böylece her sınıfa ait olasılık tahmininde bulunur.

Lojistik regresyonun matematiksel temeli Denklem 3'te gösterilmektedir. Denklem 3'te görüldüğü üzere bağımsız değişkenlerle hedef sınıf arasındaki ilişki şu şekilde tanımlanmaktadır (Hosmer Jr et al., 2013):

$$P(y = 1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}} \qquad (3)$$

Burada:

- $P(y = 1|x)$, bağımsız değişkenler (X) verildiğinde, sınıf y=1 (örneğin pozitif sınıf) olma olasılığıdır.
- $\beta_0, \beta_1, \ldots, \beta_n$ modelin katsayıları olup her bağımsız değişkenin hedef sınıf üzerindeki etkisini temsil eder.
- $X_0, X_1, \ldots, X_n$ ise giriş özellikleridir.

Algoritmanın amacı, verilerden elde edilen gözlemler doğrultusunda bu katsayıları öğrenmek ve en iyi tahminleri yapmaktır. Lojistik regresyon, doğrusal regresyonun aksine çıktı değerini sürekli bir sayı yerine, sınıflar arasında olasılıklı bir dağılım ile sunar. Bu sayede model, ikili sınıflandırma problemlerinde yaygın bir şekilde tercih edilmektedir.

### *2.4.3. Destek Vektör Makinesi (DVM)*

Destek Vektör Makinesi (DVM), verileri sınıflarına ayırmak için kullanılan güçlü bir denetimli öğrenme algoritmasıdır. DVM, sınıflandırma işlemini gerçekleştirirken verileri temsil eden noktalar arasında en iyi ayrım çizgisini (hiper düzlemi) bulmaya çalışır. Bu hiper düzlem sınıfları en doğru şekilde ayıran ve sınıflar arasındaki en büyük marjı sağlayan bir doğrusal fonksiyon ile tanımlanır. DVM'nin temel amacı, iki sınıf arasında mümkün olan en geniş marjı oluşturan hiper düzlemi bulmaktır. Bu marjı maksimize eden hiper düzlem, sınıfların daha iyi ayrılmasına olanak tanır ve yeni veri noktalarının doğru bir şekilde sınıflandırılmasını sağlar (Nurkholis et al., 2022).

Genel olarak çalışmada olduğu gibi iki sınıftan oluşan sınıflandırma problemlerinde kullanılan yöntemin ikiden fazla sınıflı problemler için geliştirilmiş farklı yaklaşımları da vardır. Bu yaklaşımlar, birden fazla sınıfın bulunduğu durumlarda sınıflar arasındaki ilişkiyi belirlemek için geliştirilmiştir (Bishop & Nasrabadi, 2006; Hsu & Lin, 2002). İkiden fazla sınıflı problemler için en yaygın kullanılan iki temel yöntem şunlardır:

- Birine Karşı Diğerleri (One-vs-Rest, OvR) Yaklaşımı: Bu yöntemde her bir sınıf için ayrı bir DVM modeli oluşturulur. Her model, bir sınıfı pozitif sınıf olarak kabul ederken diğer tüm sınıfları negatif sınıf olarak kabul eder. Kısacası k sınıflı bir problem için k adet DVM modeli eğitilir. Her yeni veri noktası, en yüksek güven skoruna sahip olan sınıfa atanır. Bu yaklaşım, çok sınıflı problemlerde yaygın olarak kullanılan uygulaması kolay bir yöntemdir.

- Çiftler Arası Karşılaştırma (One-vs-One, OvO) Yaklaşımı: Bu yöntemde her sınıf çifti için ayrı bir DVM modeli eğitilir. Örneğin k sınıflı bir problem için her iki sınıf arasında bir DVM modeli oluşturularak toplamda $k(k-1)/2$ adet DVM modeli elde edilir. Yeni bir veri noktası sınıflandırılırken, tüm bu modeller kullanılır ve en fazla oy alan sınıf son karar olarak atanır. Bu yöntem çok sayıda sınıf olduğunda daha fazla model gerektirse de genellikle daha hassas sonuçlar verebilir.

## 2.5. Performans Metrikleri

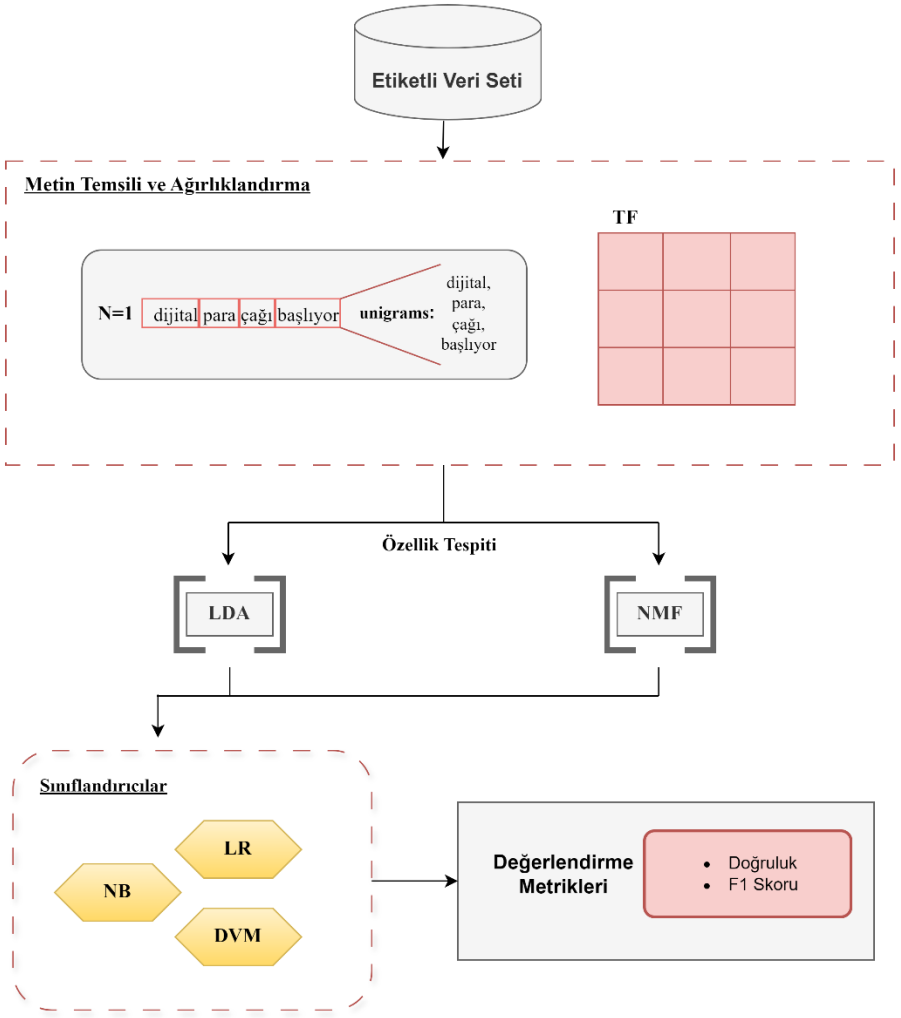
Çalışmada uygulanan sınıflandırma algoritmalarının performanslarını ölçmek için karmaşıklık matrisinde bulunan terimler kullanılarak sınıflandırma doğruluğu ve f1-ölçütü metrikleri hesaplanmıştır. Karmaşıklık matrisi doğru veya yanlış sınıflandırmaların sayısını gösteren basit ve sezgisel bir metriktir. Tablo 3 karmaşıklık matrisinde bulunan True Positive (TP), False Positive (FP), True Negative (TN) ve False Negative (FN) olmak üzere dört temel bileşenini açıklamaktadır. Tablo 4'te ise karmaşıklık matrisinden elde edilen TP, FP, TN ve FN değerlerinin kullanılması ile hesaplanan sınıflandırma doğruluğu ve F1-ölçütü metriklerinin denklemleri yer almaktadır.

*Tablo 3. Karmaşıklık Matrisi Terimleri*

| | |
|---|---|
| **True Pozitif (TP)** | Veri setinde olumlu olarak etiketli tweetlerin olumlu olarak doğru tahmin edilmesi. |
| **False Positive (FP)** | Veri setinde olumsuz olarak etiketli tweetlerin olumlu olarak yanlış tahmin edilmesi. |
| **True Negative (TN)** | Veri setinde olumsuz olarak etiketli tweetlerin olumsuz olarak doğru tahmin edilmesi. |
| **False Negative (FN)** | Veri setinde olumlu olarak etiketli tweetlerin olumsuz olarak yanlış tahmin edilmesi. |

*Tablo 4. Performans Metrikleri*

| | | |
|---|---|---|
| **Doğruluk (accuracy):** | Pozitif ve negatif sınıflarda bulunan tweetlerin doğru tahmin edilme oranını hesaplar. | $\frac{TP + TN}{TP + FP + TN + FN}$ |
| **Duyarlılık (precision):** | Pozitif tahmin edilen pozitif ve negatif tweetlerden kaç tanesinin pozitif olduğunun oranını hesaplar. | $\frac{TP}{TP + FP}$ |
| **Geri çağırma (recall):** | Pozitif tweetlerden kaç tanesinin doğru tahmin edildiğini hesaplar. | $\frac{TP}{TP + FN}$ |
| **F1-ölçütü:** | Duyarlılık ve geri çağırma metriklerinin harmonik ortalamasıdır. | $\frac{2 * duyarlılık * geri\ çağırma}{duyarlılık + geri\ çağırma}$ |




*Şekil 2. Çalışmanın Genel Mimarisi*

# 3. Araştırma Sonuçları ve Tartışma

Bu çalışmada LDA ve NMF olmak üzere iki farklı konu modelleme yöntemi karşılaştırılarak, NB, LR ve DVM sınıflandırıcılarının çeşitli parametre değerleri altındaki performansları değerlendirilmektedir. Her iki modelleme yöntemi için farklı özellik boyutları ve bileşen sayılarının kombinasyonlarıyla oluşturulan modellerin performansı, sınıflandırma doğruluğu ve F1-ölçütü metrikleri ile değerlendirilmiş olup alınan sonuçlar, Tablo 5 ve Tablo 6'da detaylı olarak sunulmuştur.

LDA konu modelleme yönteminde 5 farklı konu özellik boyutu (100, 200, 300, 400, 500), 5 farklı bileşen sayısı (10, 20, 30, 40, 50) ve farklı boyutlardaki özellik sayıların kombinasyonları ile oluşturulan modellerden elde edilen performans değerleri Tablo 5'te gösterilmiştir. Tablo 5'e göre en yüksek sınıflandırma doğruluğu tüm modeller arasında DVM sınıflandırıcısı ile; konu özellik boyutu ve bileşen sayısı parametreleri için sırasıyla 500 ve 40 değerleri girildiğinde %87.89 olarak ulaşılmıştır.

Kullanılan diğer sınıflandırıcılar için en yüksek performans değerlerine bakıldığında LR sınıflandırıcı için en yüksek sınıflandırma doğruluğu değeri konu özellik boyutu 300 ve bileşen sayısı parametresi 40 olarak girildiğinde %87.14; NB sınıflandırıcı için ise konu özellik boyutu 200 ve bileşen sayısı 50 olarak girildiğinde %79.66 olarak elde edilmiştir. Bu sonuçlar, LDA modeli ile farklı sınıflandırıcıların ve parametrelerin, sınıflandırma performansı üzerindeki etkisini göstermektedir.

NMF konu modelleme yöntemi için de aynı şekilde 5 farklı konu özellik, 5 farklı bileşen sayısı ve farklı boyutlardaki özellik sayıların kombinasyonları ile oluşturulan modellerden elde edilen performans değerleri Tablo 6'da gösterilmiştir. Tablo 6'ya göre tüm modeller arasında en yüksek sınıflandırma doğruluğu DVM sınıflandırıcısı ile; konu özellik boyutu ve bileşen sayısı parametreleri için sırasıyla 500 ve 50 değerleri girildiğinde %87,43 olarak ulaşılmıştır.

Kullanılan diğer sınıflandırıcılar için en yüksek performans değerlerine bakıldığında LR sınıflandırıcı için en yüksek sınıflandırma doğruluğu değeri konu özellik boyutu 500 ve bileşen sayısı parametresi 50 olarak girildiğinde %87,02; NB sınıflandırıcı için ise konu özellik boyutu 500 ve bileşen sayısı 40 olarak girildiğinde %79.78 olarak elde edilmiştir. Bu sonuçlar da aynı şekilde NMF modellemesinde benzer parametre kombinasyonlarının sınıflandırıcı performanslarını nasıl etkilediğini ortaya koymaktadır.

*Tablo 5. LDA Yöntemi ile Oluşturulan Modellerden Elde Edilen Performans Değerleri*

| Konu Özellik Boyutu | Bileşen Sayısı | Kullanılan Özellik Sayısı | NB | | LR | | DVM | |
|---|---|---|---|---|---|---|---|---|
| | | | Sınıflandırma Doğruluğu | F1-ölçütü | Sınıflandırma Doğruluğu | F1-ölçütü | Sınıflandırma Doğruluğu | F1-ölçütü |
| 100 | 10 | 568 | 73,0096 | 73,0141 | 75,6241 | 75,2248 | 75,6961 | 75,3381 |
| | 20 | 1017 | 75,4261 | 75,6579 | 79,3139 | 79,0051 | 79,3949 | 79,1363 |
| | 30 | 1433 | 76,9784 | 77,1718 | 81,5113 | 81,2635 | 81,5983 | 81,3993 |
| | 40 | 1909 | 78,3437 | 78,6634 | 83,5617 | 83,3572 | 83,6727 | 83,5056 |
| | 50 | 2447 | 78,6777 | 79,0298 | 84,5849 | 84,4388 | 84,7690 | 84,6576 |
| 200 | 10 | 1089 | 76,2943 | 76,5015 | 80,3431 | 80,0579 | 80,4121 | 80,1805 |
| | 20 | 1883 | 77,9346 | 78,2810 | 83,2276 | 83,0297 | 83,3767 | 83,2185 |
| | 30 | 2627 | 79,0538 | 79,3929 | 85,4491 | 85,3081 | 85,5771 | 85,4554 |
| | 40 | 3428 | 79,5359 | 79,8880 | 86,3453 | 86,2275 | 86,6073 | 86,5031 |
| | 50 | 4318 | **79,6649** | **80,0206** | 86,8964 | 86,8013 | 87,2865 | 87,2003 |
| 300 | 10 | 1563 | 77,7926 | 78,0172 | 82,7095 | 82,4746 | 82,8226 | 82,6408 |
| | 20 | 2656 | 79,0278 | 79,3701 | 85,3631 | 85,2120 | 85,5141 | 85,3812 |
| | 30 | 3686 | 79,6429 | 79,9971 | 86,8364 | 86,7390 | 87,0924 | 87,0131 |
| | 40 | 4805 | 79,6059 | 79,9605 | **87,1374** | 87,0339 | 87,5525 | 87,4662 |
| | 50 | 5956 | 79,3819 | 79,7548 | 87,0734 | 86,9806 | 87,7395 | 87,6551 |
| 400 | 10 | 2016 | 78,8748 | 79,1020 | 84,5729 | 84,4079 | 84,6709 | 84,5393 |
| | 20 | 2016 | 79,5569 | 79,8852 | 86,4953 | 86,3783 | 86,7594 | 86,6535 |
| | 30 | 4668 | 79,5449 | 79,8983 | 87,0604 | 86,9656 | 87,5295 | 87,4547 |
| | 40 | 4068 | 79,3459 | 79,6919 | 87,1324 | 87,0423 | 87,7686 | 87,6892 |
| | 50 | 7497 | 79,0948 | 79,4752 | 86,9564 | 86,8805 | 87,8616 | 87,7887 |
| 500 | 10 | 2475 | 79,3569 | 79,6317 | 85,6621 | 85,5149 | 85,7562 | 85,6279 |
| | 20 | 4150 | 79,6179 | 79,9703 | 86,9614 | 86,8767 | 87,3695 | 87,2874 |
| | 30 | 5669 | 79,4359 | 79,7964 | 87,1334 | **87,0538** | 87,7405 | 87,6666 |
| | 40 | 7316 | 79,0408 | 79,4073 | 87,0834 | 87,0063 | **87,8856** | **87,8117** |
| | 50 | 9071 | 78,7898 | 79,1959 | 86,7103 | 86,6403 | 87,8576 | 87,7858 |

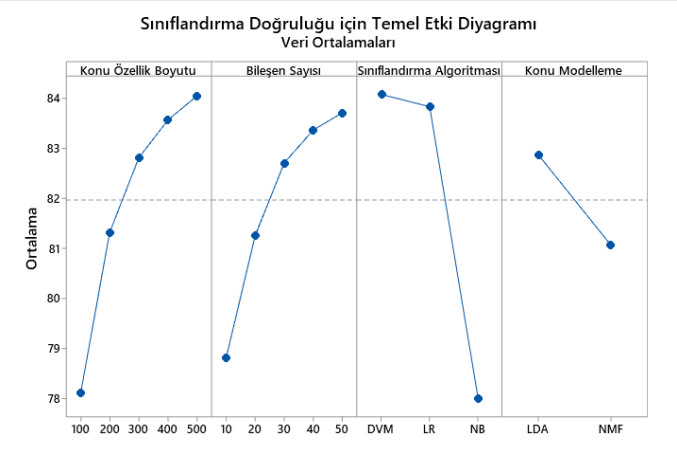
*Tablo 6. NMF Yöntemi ile Oluşturulan Modellerden Elde Edilen Performans Değerleri*

| Konu Özellik Boyutu | Bileşen Sayısı | Kullanılan Özellik Sayısı | NB | | LR | | DVM | |
|---|---|---|---|---|---|---|---|---|
| | | | Sınıflandırma Doğruluğu | F1-ölçütü | Sınıflandırma Doğruluğu | F1-ölçütü | Sınıflandırma Doğruluğu | F1-ölçütü |
| 100 | 10 | 421 | 70,0050 | 70,0929 | 72,3764 | 71,3681 | 72,4184 | 71,5509 |
| | 20 | 674 | 73,2196 | 73,3531 | 76,2512 | 75,6905 | 76,3242 | 75,8527 |
| | 30 | 992 | 75,3350 | 75,4452 | 79,2388 | 78,8167 | 79,3158 | 78,9655 |
| | 40 | 1189 | 76,1222 | 76,2788 | 80,4690 | 80,1123 | 80,4970 | 80,2026 |
| | 50 | 1392 | 76,7223 | 76,8957 | 81,2942 | 80,9677 | 81,3842 | 81,1198 |
| 200 | 10 | 745 | 73,6987 | 73,9057 | 76,8523 | 76,3323 | 76,8993 | 76,4795 |
| | 20 | 1168 | 76,1872 | 76,4230 | 80,3520 | 79,9949 | 80,4080 | 80,1096 |
| | 30 | 1656 | 77,5705 | 77,8455 | 82,6665 | 82,3896 | 82,7555 | 82,5297 |
| | 40 | 1988 | 78,2136 | 78,5264 | 83,7727 | 83,5684 | 83,8947 | 83,7348 |
| | 50 | 2285 | 78,7817 | 79,0952 | 84,7359 | 84,5397 | 84,8319 | 84,6704 |
| 300 | 10 | 1068 | 75,5501 | 75,7506 | 79,6289 | 79,2388 | 79,6519 | 79,3385 |
| | 20 | 1640 | 77,4964 | 77,8013 | 82,6125 | 82,3415 | 82,7415 | 82,5185 |
| | 30 | 2251 | 78,7137 | 79,0027 | 84,5379 | 84,3414 | 84,7339 | 84,5715 |
| | 40 | 2691 | 79,3448 | 79,6568 | 85,6571 | 85,4868 | 85,8811 | 85,7482 |
| | 50 | 3112 | 79,6269 | 79,9608 | 86,2942 | 86,1540 | 86,5283 | 86,4128 |
| 400 | 10 | 1363 | 76,7393 | 77,0096 | 81,3032 | 80,9596 | 81,3682 | 81,0840 |
| | 20 | 2083 | 78,3696 | 78,6986 | 84,1238 | 83,9161 | 84,2588 | 84,0918 |
| | 30 | 2823 | 79,4708 | 79,7497 | 85,7611 | 85,5970 | 85,9941 | 85,8601 |
| | 40 | 3352 | 79,6759 | 79,9887 | 86,4112 | 86,2737 | 86,6553 | 86,5400 |
| | 50 | 3906 | 79,7129 | 80,0534 | 86,7933 | 86,6699 | 87,1094 | 87,0056 |
| 500 | 10 | 1672 | 77,8845 | 78,1782 | 83,0506 | 82,7815 | 83,1056 | 82,8934 |
| | 20 | 2499 | 79,1978 | 79,4899 | 85,3080 | 85,1378 | 85,4300 | 85,2865 |
| | 30 | 3412 | 79,7069 | 80,0113 | 86,5113 | 86,3705 | 86,7753 | 86,6584 |
| | 40 | 4069 | **79,7779** | **80,1037** | 86,8273 | 86,7108 | 87,1834 | 87,0776 |
| | 50 | 4739 | 79,7559 | 80,1009 | **87,0174** | **86,9087** | **87,4384** | **87,3423** |

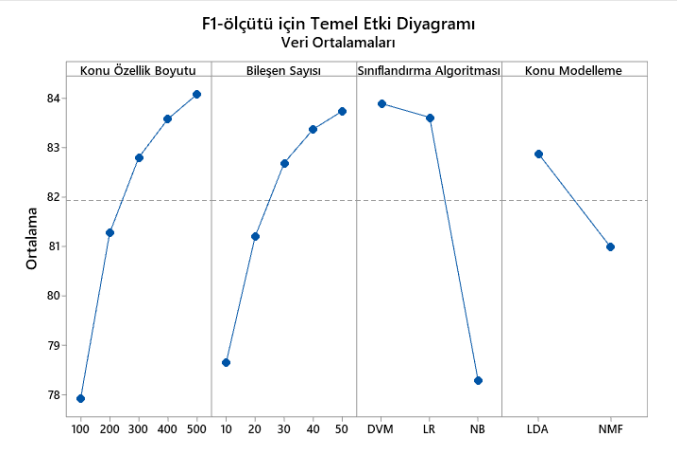
Deneysel analizde farklı konu özellik boyutu, bileşen, sınıflandırıcı ve konu modelleme yöntemleri kullanıldığında sınıflandırma doğruluğu ve f1-ölçütü metrikleri için elde edilen ortalama değerler Şekil 3 ve Şekil 4'te istatistiksel olarak gösterilmiştir.

- Şekil 3 ve Şekil 4'te görüldüğü üzere konu özellik boyutu için girilen değerler arttırıldığında hem sınıflandırma doğruluğu hem de F1-ölçütü değerlerinde anlamlı bir artış gözlenmektedir. Bu durum daha fazla özellik kullanıldığında modellerin daha iyi bir şekilde eğitildiğini göstermektedir.
- Benzer şekilde, bileşen sayısındaki artış hem sınıflandırma doğruluğu hem de F1-ölçütü için pozitif bir etkiye sahiptir.
- Sınıflandırıcılar arasında ortalama değer için en yüksek sonuçlara, iki performans değerlendirme metriğinde de LR sınıflandırıcı ile ulaşılmış olup DVM ile de en yüksek ikinci en iyi ortalama performans değerleri elde edilmiştir.
- NB ise snıflandırıcılar arasında en kötü performansı göstermiştir.
- Son olarak konu modelleme yöntemleri arasında ortalama değerlere göre kıyaslama yapıldığında ise her iki grafikte de LDA yöntemi ile NMF yöntemine göre daha yüksek sonuçlar alınmıştır.

Bu sonuçlar, model optimizasyonu için konu özellik boyutu ve bileşen sayısının artırılmasının ve uygun sınıflandırıcı ile çalışılmasının performans artışındaki öenmini göstermektedir.




*Şekil 3. Sınıflandırma Doğruluğu Metriği Için Elde Edilen Ortalama Değerler*




*Şekil 4. F1-Ölçütü Metriği Için Elde Edilen Ortalama Değerler*

# 4. Sonuç

2008 yılında yayınlanan bir makale ile adını duyuran ve 2009 yılında satışa sunulan ilk ve en büyük kripto para birimi Bitcoin ile hayatımıza giren kripto para kavramı, insanlar tarafından oldukça merak edilen ve araştırılan bir yeniliktir. Bitcoin, ilk ve en büyük kripto para birimi olarak finansal sistemlerde önemli bir devrim yaratmış ve günümüzde birçok farklı kripto para biriminin oluşmasına öncülük etmiştir. Kripto paralar gün geçtikçe daha çok insan tarafından benimsenmekte ve yatırım aracı olarak kullanılmaktadır.

Son yıllarda popüler sosyal medya platformu olan Twitter üzerinden yapılan kripto paralar ile ilgili paylaşımların sayısı oldukça artmış ve platform kripto paralar ile ilgili büyük bir veri kaynağı haline gelmiştir. Twitter'ın hızlı, anlık ve geniş kitlelere ulaşabilen yapısı, kripto paralar hakkında büyük bir veri kaynağı sunmakta ve yatırımcılar tarafından piyasa hareketlerinin değerlendirilmesinde kullanılmaktadır. Bu çalışmada konu ile ilgilenen insanlar için bir duygu analizi çalışması yaparak piyasanın yönü hakkında bilgi vermek ve literatüre katkı sağlamak hedeflenmiştir. Bu amaç doğrultusunda, Twitter platformu üzerinden konu ile ilgili paylaşılan tweetler çekilmiş ve LSTM ile otomatik etiketleme işlemi yapılarak bu çalışmada kullanılacak yeni bir veri seti oluşturulmuştur. LSTM, özellikle zaman serisi verilerinde etkili olan bir derin öğrenme modelidir ve dilsel veri üzerinde güçlü performans göstermesi nedeniyle bu çalışmada tercih edilmiştir. Bu şekilde oluşturulan yeni veri setinin daha önceki çalışmalara kıyasla daha doğru ve güncel sonuçlar elde edilmesine olanak tanıdığı düşünülmektedir.

Çalışmada kullanılacak veri seti hazırlandıktan sonra, metinlerden elde edilen özelliklerin çıkarılması için TF ağırlıklandırma ve n-gram kelime temsil yöntemleri kullanılmıştır. TF ağırlıklandırma, kelimelerin metinlerdeki frekanslarını hesaplayarak önemli kelimelerin belirlenmesine yardımcı olurken, n-gram modeli ise kelimelerin ardışık dizilimlerini inceleyerek bağlamsal ilişkilerin yakalanmasını sağlamaktadır. Özellikle çalışmada kullanılan veri setinde olduğu gibi belirli bir konuya odaklanan metinlerde, bu tür kelime temsil yöntemleri daha anlamlı sonuçlar üretmektedir.

Özellik tespiti için LDA ve NMF konu modelleme yöntemleri tercih edilmiştir. Her iki yöntem de kripto paralarla

ilgili tweetlerdeki temel eğilimlerin ve kullanıcıların ilgi alanlarının belirlenmesine katkı sağlamıştır.

Farklı parametre değerleri ve NB, LR ve DVM olmak üzere 3 farklı sınıflandırıcı kullanılarak oluşturulan modeller arasında en yüksek performans değeri LDA yöntemi ve DVM sınıflandırıcı kullanıldığında %87.89 olarak elde edilmiştir. Bu sonuç, konu modelleme ve sınıflandırma süreçlerinin doğru parametreler ve modeller ile kullanıldığında kripto para piyasası hakkında anlamlı tahminler yapma konusunda etkili olduğunu göstermektedir.

# 5. Öneriler

Bu çalışma, kripto para piyasası üzerine yapılan duygu analizi çalışmalarına önemli bir katkı sağlamakta olup ileride yapılacak çalışmalar için de çeşitli öneriler sunmaktadır. Aşağıda gelecekte yapılabilecek araştırmalar için sunulan öneriler sıralanmıştır:

1. **Veri Setinin Genişletilmesi:** Mevcut veri seti genişletilerek farklı zaman aralıkları ve daha fazla kripto para birimi üzerine yoğunlaşan tweetler eklenebilir. Daha fazla veri sayısıyla analiz yapmak sonuçların genelliğini artırabilir.
2. **Veri Setindeki Sınıf Sayısının Artırılması:** Mevcut çalışmada kullanılan veri seti, daha fazla sınıfa ayrılabilir. Bu sayede sadece olumlu, olumsuz ve nötr gibi genel sınıflar yerine heyecan, korku, endişe, memnuniyet gibi duygu kategorileri oluşturularak duyguların piyasa hareketleri üzerindeki etkilerinin daha derinlemesine incelenmesi mümkün olabilir.
3. **Farklı Sınıflandırma Algoritmalarının Kullanılması:** Mevcut sınıflandırıcılara ek olarak farklı algoritmaların kullanılması ve performanslarının karşılaştırılması sonuçların iyileştirilmesine katkı sağlayabilir.
4. **Derin Öğrenme Modellerinin Geliştirilmesi:** Daha karmaşık modellerin ve derin öğrenme tekniklerinin kullanılmasıyla kripto paralarla ilgili daha detaylı ve doğru öngörülerde bulunulması mümkündür.

Sıralanan önerilerin, kripto para piyasasının daha iyi anlaşılmasına ve yatırımcıların daha bilinçli kararlar almasına yardımcı olacağı düşünülmektedir.

**Kaynakça**


Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems*, *112*, 102131.

Altıntaş, V., Albayrak, M., & Topal, K. (2021). Kanser Hastalığı İle İlgili Paylaşımlar İçin Dirichlet Ayrımı İle Gizli Konu Modelleme. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, *36*(4), 2183-2196.

Ankaralı, E., & Külcü, Ö. (2020). RapidMiner ile Twitter Verilerinin Konu Modellemesi. *Bilgi Yönetimi*, *3*(1), 1-10.

Aydın, G., & Hallaç, İ. (2021). Türkçe Metinlerde Otomatik Konu Tespiti. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, *33*(2), 599-606.

Balli, S., & Sağbas, E. A. (2017). *Advances in Statistical Methodologies and Their Application to Real Problems* (T. Hokimoto, Ed.). https://doi.org/10.5772/62963

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Böttger, J. (2017). Understanding benefits of different vantage points in today's Internet.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *20*(2), 215-232.

Çoban, Ö., & Tümüklü Özyer, G. (2018). Twitter duygu analizinde terim ağırlıklandırma yönteminin etkisi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, *24*(2), 283-291.

Das, B., & Chakraborty, S. (2018). An improved text sentiment classification model using TF-IDF and next word negation. *arXiv preprint arXiv:1806.06407*.

Dewi, C., Chen, R. C., Christanto, H. J., & Cauteruccio, F. (2023). Multinomial Naive Bayes Classifier for Sentiment Analysis of Internet Movie Database. *Vietnam Journal of Computer Science*.

Ekinci, E., & Omurca, S. İ. (2017). Ürün özelliklerinin konu modelleme yöntemi ile çıkartılması. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *9*(1), 51-58.

Günay, K., & Güçdemir, Y. (2020). İklim İletişimi Bağlamında 2020-2021 Türkiye'de Sivil Toplum Kuruluşların Twitter Paylaşımlarının Konu Modelleme Analizi. *Turkish Online Journal of Design Art and Communication*, *12*(4).

Habek, G. C., Toçoğlu, M. A., & Onan, A. (2022). *Farklı Kripto Para Çeşitleri ile Duygu Analizi* 1st International Conference on Scientific and Academic

Research (ICSAR), Konya, Turkey.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, *13*(2), 415-425.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Koruyan, K. (2022). BERTopic Konu Modelleme Tekniği Kullanılarak Müşteri Şikayetlerinin Sınıflandırılması. *İzmir Sosyal Bilimler Dergisi*, *4*(2), 66-79.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *nature*, *401*(6755), 788-791.

M'sik, B., & Casablanca, B. M. (2020). Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. *International Journal*, *9*(4).

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. AAAI-98 workshop on learning for text categorization,

Nurkholis, A., Alita, D., & Munandar, A. (2022). Comparison of Kernel Support Vector Machine Multi-Class in PPKM Sentiment Analysis on Twitter. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(2), 227-233.

Onan, A., & Bayar, C. (2017). Türkçe Twitter Mesajlarında Gizli Dirichlet Tahsisine Dayalı Duygu Analizi. *Akademik Bilişim*, *8*(10).

Onan, A., Yalçın, A., & Atik, E. (2020). Üniversite Bilgi Yönetim Sistemi Servis Destek Taleplerinin Konu Modelleme Tabanlı Analizi. *Avrupa Bilim ve Teknoloji Dergisi, Ejosat Özel Sayı 2020 (HORA)*, 389-397.

Shoetan, P. O., & Familoni, B. T. (2024). Blockchain's impact on financial security and efficiency beyond cryptocurrency uses. *International Journal of Management & Entrepreneurship Research*, *6*(4), 1211-1235.

Turan, S. C., Yildiz, K., & Büyüktanir, B. (2023). Comparison of LDA, NMF and BERTopic Topic Modeling Techniques on Amazon Product Review Dataset: A Case Study. International Conference on Computing, Intelligence and Data Analytics,

Wang, J., & Zhang, X. L. . (2023). Deep NMF Topic Modeling. *Neurocomputing*(515), 157-173.

Zhang, H. (2004). The optimality of naive Bayes. *Aa*, *1*(2), 3.