Sosyal Medyada Kripto-Para Duyarlılık Analizi
Crypto-Currency Sentiment Analyse on Social Media

Mehmet Can ERDOĞAN
Van Yüzüncü Yıl Üniversitesi,
Başkale MYO
canerdogan@yyu.edu.tr

Murat CANAYAZ
Van Yüzüncü Yıl Üniversitesi,
Bilgisayar Mühendisliği Bölümü
mcanayaz@yyu.edu.tr

***Özet*— Sosyal medya araçlarının hayatımıza girmesi ile üretilen veri miktarı baş döndürücü boyutlara ulaşmıştır. Verileri analiz etmekte, geleneksel yöntemlerin artık yetersiz geldiği günümüzde "Büyük Veri" kavramı hayatımıza girmiştir. Devasa boyuttaki verileri analiz ederek anlamlı özetler çıkarmak kaçınılmaz bir ihtiyaç olmaktadır. Bu ihtiyacı karşılamak için büyük veri araçları kullanılmaktadır. Bu araçları kullanarak insan davranışları hakkında bilgi sahibi olmak ve bu doğrultuda çözümler geliştirmek için, sosyal medya ve özellikle Twitter verileri üzerinde çalışmalar yapılmaktadır. Bilindiği üzere, son yılların trend konularından olan kripto para kavramı, daha fazla insanın ilgisini çekmektedir. Bu çalışmada, büyük veri araçları kullanılarak, en çok kullanılan kripto para birimlerine olan yaklaşımlar ilk defa incelenmiştir. Twitter'dan elde edilen twitler üzerinde anlamsız veriler temizlenmiş, kullanıcıların kripto paralara karşı yaklaşımları çeşitli sınıflandırıcılar kullanılarak analiz edilmiş ve sonuçlar gösterilmiştir.**

Anahtar kelimeler: Büyük veri, sosyal medya, Twitter, kripto para, metin sınıflandırma, duyarlılık analizi

Abstract -The amount of data produced with the introduction of social media tools has reached gigantic size. In analyzing data, and now traditional methods are no longer enough, the concept of "Big Data" has entered our lives. There is an inevitable need to produce meaningful summaries by analyzing data at huge size. Large data tools are used to meet this need. In order to have knowledge about human behaviors using these tools and to develop solutions in this direction, social media and especially Twitter data are being studied. As is known, the concept of crypto currency, which has come from trends in recent years, attracts more and more people. In this study, approaches to the most commonly used crypto currencies were examined for the first time using large data tools. The meaningless data on twits which is taken from Twitter has been cleared and the approaches of users against crypto paralysis have been analyzed by using various classifiers and the results have been shown.

Keyword: Big data, social media, Twitter, crypto currency, text classification, sentiment analyse

I. GİRİŞ

Günümüzde teknolojinin gelişmesi, internetin daha yaygın kullanımı ve gelişen sosyal medya araçları sayesinde üretilen bilgi miktarı baş döndürücü hızlara ulaşmıştır.

Gelişen internet ile beraber hayatımıza yeni terimler girmektedir ve girmeye devam edecektir. Artık sadece bilgisayarlar ve telefonlar değil birçok cihaz internete bağlanabilmektedir. IoT (Internet of Things-Nesnelerin İnterneti), bilgisayar ve telefonlar dışındaki birçok elektronik cihaz internete erişerek birbirleriyle ya da daha büyük sistemlerle iletişimde olduğu bir ağdır. Mutfak gereçleri, güvenlik sistemleri, kameralar, fotoğraf makineleri, arabalar bu ağa bağlanmaya başlamış ve önümüzdeki birkaç yıl içinde daha fazla cihaz bu listeye eklenecektir.

İnternete bağlı bu kadar çok sayıda cihaz varken üretilen veri miktarı hakkında fikir sahibi olabilmek için birçok araştırma şirketi, verinin büyüklüğü ile ilgili şimdiye kadar üretilen ve bundan sonra ulaşacağı boyutlar ile alakalı tahminlerde bulunmaktadır.

IDC (International Data Corporation), büyük veriler üzerinde araştırmalar yapan ve düzenli raporlar sunan bir şirkettir. IDC verilerine göre, 2005 yılında 130 exabyte iken 2020 yılında 40 bin exabyte veya 40 trilyon gigabyte olması öngörülmektedir. Bu, kişi başına 5 terrabyte'ın üzerindeki bir rakama tekabül etmektedir. Şu andan itibaren 2020 yılına kadar, internetteki veri miktarı, her iki yılda bir, yaklaşık iki katına çıkacak. IDC 2020 yılında, dijital evrenin bugünkü % 25 ile karşılaştırıldığında, %33'nün analiz edildiği takdirde değerli olabilecek bilgiler içereceğini tahmin etmektedir[1].

Günümüzde toplumu bilgi toplumu olarak adlandırabilir ve bunun kanıtlarını her yerde çok kolay görebiliriz. İnsanların neredeyse tamamında akıllı telefon, hemen her evde bilgisayar veya tablet ve tüm şirket ve kurumların bilgi teknolojileri birimi bulunmaktadır. Ancak bütün bilgiler okunur durumda veya anlamlı değildir. Günümüzde sadece bilgi miktarının artmasının yanı sıra bilgiye erişim hızında ve sayısında da artış gözlenmektedir. Verinin anlamlı bir bütün oluşturacak şekilde toplanması ilk önce astronomi ve genetik alanında gerçekleşmiştir. Büyük veri kavramı da ilk olarak bu alanlarda kullanılmış, ancak daha sonra birçok alanda verilerin artması ile bu kavram kullanılmaya başlanmıştır. Büyük veri artık hayatımızın neredeyse her alanında kendini göstermeye başlamıştır [2].

Veriler bu kadar artarken, bütün veriler anlamlı veya yorumlanabilir durumda değildir. Bu verileri analiz etmek ve

978-1-5386-6878-8/18/$31.00 ©2018 IEEE

faydalı bilgiyi içinden alabilmek için “Büyük Veri” ve “Veri Madenciliği” kavramları hayatımıza girmiştir. Gartnet Group’a göre, günümüzde sıkça adını duymaya başladığımız Big Data (Büyük Veri), her ne kadar teknolojinin ilerlemesi ve kullanım alanlarının artması ile ortaya çıkmış bir kelime olarak görülse de, yıllardır içerisinde bulunduğumuz fakat gelişiminden pek haberdar olamadığımız bir olgudur. Bu olguya farkında olmadan sürekli destek vermekte ve “Büyük Veri” olarak isimlendirdiğimiz bu ortama sürekli veri akışı sağlanmasında bizler de katkı sağlamaktayız [3].

2000’li yıllarda büyük veri üç bileşenli olarak tanımlanmış ve İngilizce baş harflerinden dolayı 3V olarak isimlendirilmiştir [4].

Volume (Hacim veya veri büyüklüğü): Üretilen verinin miktarı ve saklanan verinin boyutunu temsil eder. Üretilen veri miktarında göre o verinin aslında büyük veri olup olmadığına karar verilir. Veriyi algılayıcılar, süper bilgisayarlar, kişisel bilgisayarlar, sunuculara, arabalar, uçaklar gibi birçok farklı kaynak üretmektedir.

Velocity (Hız): Büyük veri, yukarı doğru ivmeli bir hızla üretilmekte ve bu veriler saniyede inanılmaz boyutlara ulaşmaktadır. Hızlı büyüyen veri, o veriyi kullanan işlem sayısının ve çeşitliliği de aynı hızda artmaktadır ve hem yazılım hem de donanım olarak bu yoğunluğa cevap verilebilmektedir.

Variety (Çeşitlilik): Üretilen veriler genel olarak düzenli olmadığı ve birçok farklı ortamdan elde edilen veri biçimlerinden oluştukları için, verilerin sistemli ve birbirlerine dönüştürülebiliyor olmaları gerekmektedir.

- Dijital verilerin %70-%80 düzenli olmayan veriler oluşturur.
- Anlamlı bilgilerin %80-%90 düzenli olmayan verilerden elde edilir [5].
- Veri madenciliği, Doğal Dil işleme, makine öğrenmesi gibi alanlar bu verileri yorumlamaya çalışmaktadır.

2000’li yıllarda büyük veriye yaklaşım 3V olarak kabul edilirken günümüzde büyük veri ve büyük veriye yaklaşım değişkenlik gösterip artık 5V ile gösterilmektedir [6,7].

Verification (Doğrulama): Veri üretiminin bu kadar hızlı olan günümüzde bu verilerin güvenli olup olmadığı önemlidir. Bunun için dördüncü V olan Verification (Doğrulama), verilerin doğruluğunu kontrol etme aşamasıdır.

Value (Değer): Belki de en önemli katmanlardan bir tanesi de “Değer” katmanıdır, verilerimiz yukarıdaki veri bileşenlerinden filtrelendikten sonra büyük verinin üretimi ve işlenmesi katmanlarında elde edilen verilerin şirketler için artı değer sağlıyor olması gerekiyor.

II. YAPILAN ÇALIŞMALAR

İnternet ortamındaki verilerin bu kadar büyümesindeki en büyük etken, günümüzde sosyal medya olarak ön plana çıkmaktadır.

Büyük veri, web sayfalarının sunucu günlükleri(web server log), internet günlükleri (bloglar), cep telefonları, iletişim kayıtları gibi çeşitli kaynaklardan gelen çok miktardaki bilgiyi içermektedir [8].

Günümüzde internetin yaygınlaşması ile birlikte, özellikle sosyal medya üzerinden çok büyük miktarda bilgi paylaşmaktadırlar. Bu verilerin doğru analiz edilmesi ve doğru metotlarla yorumlanması, değerli bilginin içinden seçilmesi önem kazanmaktadır. Bu verileri anlamlı hale getirebilen işletmeler, risklerini daha iyi yönetebilmekte, yenilik yapabilmekte ve pazarlama stratejisi oluşturabilmektedirler. Şirketler, bir adım öne geçebilmek amacıyla, iş yapma şekillerinin değiştiği çağımızda, fark yaratmak zorundadırlar[9].

Büyük verileri analiz edip anlamlı ve değerli bilgiyi içinden süzmek için veri madenciliği yapılmaktadır. Bu kavram hayatımıza daha yeni girmiş, konuyla ilgili işin uzmanları daha yeni yetişmektedir.

Veri madenciliği; önceden dağınık ve anlamsız olan verilerin, geçerli ve uygulanabilir veri haline getirme süreci olarak tanımlanabilir. Bu süreçte veri özetleme, sınıflandırma, bağımlılık ağlarının bulunması, değişkenlik analizi gibi farklı birçok teknik kullanılmaktadır[10].

Veri madenciliği, büyük verinin analiz edilerek anlamlı ve değerli bilgiyi çok geniş bir bilgi kalabalığından çekmeye çalışmaktır. Bu nedenle veri tabanları ile yakından ilişkilidir. Veri madenciliği, aynı zamanda, gerekli bilgileri elde ederek problemleri çözmek için karar verme sürecini destekleyen bir araçtır. Veri madenciliği, veriler arasındaki ilişkiyi ve anlamlı bütünlüğü anlamaya çalışmaktır [11].

Günümüzde daha çok insanın internete ulaşması ve sosyal medyayı kullanımlarının artmasına paralel olarak üretilen veriler de büyümektedir ve önemi de aynı ölçüde artmaktadır. Büyük veri geldiği durumu ve günümüzde analizinin önemini birçok örnek ile açıklayabiliriz.

Data Never Sleep ismi ile grafik şeklinde istatistiksel veri paylaşan Domo firmasına göre her bir dakikada [12];

- Twitter’da 9678 tweet atılmaktadır.
- Youtube’a 400 saatlik video yüklenmektedir.
- Instagram’da kullanıcılar tarafından 2.430.555 beğeni yapılmaktadır.
- Snapchat’ta 7 milyona yakın video izlenme sayısı oluşmaktadır.
- Depolama alanı olan Dropbox’a 8.333.333 yeni dosya yüklenmektedir.

Bu boyutlardaki veriyi analiz etmek çok her yönüyle çok önemlidir. Pazarlama, teknoloji, gelir yönetimi, güvenlik, suçların önlenmesi gibi birçok alanda kullanılmaktadır.

Amerika’nın Seattle ve Los Angeles eyaletlerinde "Önleyici Polis Hizmetleri" adında uygulama hayata geçirilmiştir. İşlenen suçların zamanı, yeri ve içeriği incelenmiştir. 4 aylık bir zaman diliminde cinayet oranını yüzde 12 gibi bir oranda düştüğü gözlenmiştir. Hırsızlık, yüzde 26 azalmıştır [13].

Hollanda istatistik birimi, yollarda bulunan sensörlerden toplanan verileri toplayarak yolların kullanım oranlarını tespit etmiştir. Sensörler önünden geçen her aracın hızını ve tipini (araba, kamyon ve benzeri) algılayarak merkezi sisteme

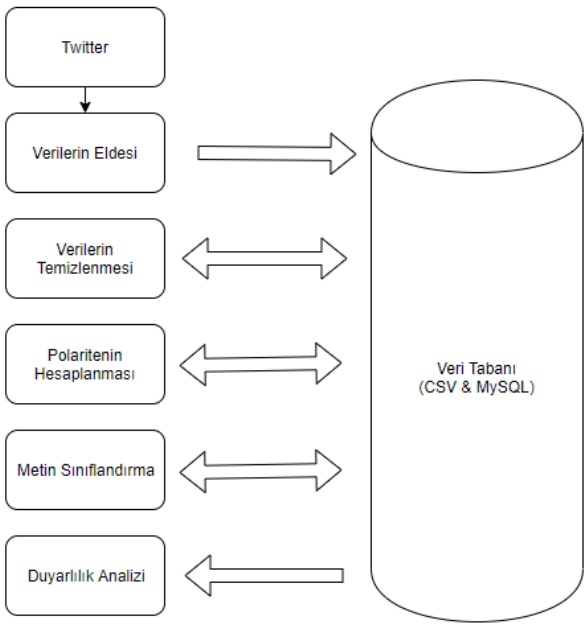
bildirmektedir. Bu çalışmayla ulaşımda alınması gereken önlemler belirlenmiştir. Projede toplanan veri boyutunun yüksekliği nedeniyle büyük veri teknolojileri tercih edilmiştir [14].

PriceStats, çevrimiçi fiyatlardan faydalanarak enflasyon tahmini yapan bir yazılım olup bu yazılım, doktora öğrencisi Alberto Cavallo tarafından geliştirilen şuan ise geniş kapsamlı bir projedir [15]. 2012 yılında "The Economist" dergisi, Arjantin devletinin resmi enflasyon rakamları yerine bu proje kapsamında hesaplanan enflasyon verilerini kullanmaya başlamıştır. Ekonomistler, PriceStats'ın büyük veriyi kullanarak hesapladığı rakamların, ülkelerin resmi istatistiklerinden daha güvenilir bulmaktadır. 2014 yılında ülkemiz de proje kapsamına alınmıştır.

Günümüzde bu kadar önemli hale gelen büyük veriler için, ülkeler ciddi bütçeler ayırmaya başlamıştır. Gartner'ın 2013 raporuna göre ABD'deki şirketlerin yüzde 64'ü ya Büyük Veri'ye yatırım yapıyor veya yapmayı planlıyor.

III. Önerilen Yöntem ve Deneysel Sonuçlar

Bu bölümde, verilerin nasıl elde edildiği ve bu veriler üzerinde yapılan işlemler ele alınmıştır. Şekil 1'de yapılan işlemlerin aşamaları gösterilmiştir.



Şekil 1. Duyarlılık Analizi Adımları

A. Verilerin Eldesi

Çalışma kapsamında en çok kullanılan beş kripto para birimi (bitcoin, ethereum, bitcoin cash, litecoin, ripple) ile ilgili atılan tweetlerin eldesi hedeflenmiştir. Bunun için Python programlama dili ile Twitter Streaming API'sini kullanan bir program geliştirilmiştir. Twitter Streaming API'sine gerekli kayıt işlemi yapıldıktan sonra konu ile ilgili olabilecek anahtar kelimeler belirlenmiş ve ilgili tweetleri, program dışarıdan durdurulana kadar dinleyerek kaydedecek şekilde programlanmıştır. Her bir para birimine ait 50 bin civarı tweet elde edilerek CSV dosyasına ve MySQL veritabanına kaydedilmiştir.

B. Verilerin Temizlenmesi

Elde edilen tweet metinleri içinde gereksiz görülen karakterler, linkler, fotoğraflar temizlenmiştir. Tablo 1'dekaydedilen tweet örneği gösterilmiştir.

TABLO 1: Elde Edilen Verinin İçeriği

tweet_id	973281594747138056
username	jylz
created_at	Mon Mar 12 19:36:29 +0000 2018
tweet	RT @giveawaysBTC: Daily Monday giveaway! 24 hour .05btc giveaway. To enter retweet, like, and follow. #btc #bitcoin #crypto #cryptocurre
cleaned_tweet	RT Daily Monday giveaway 24 hour 05btc giveaway To enter retweet like and follow btc bitcoin crypto cryptocurre
retweeted	False
user_location	Los Angeles, CA
polarity	0
group	Bitcoin

C. Polaritenin Hesaplanması

Önceden oluşturulmuş yaklaşık 3000 kelime ve her kelime için kullanılabilecek bilgilerin bulunduğu bir XML dosyasından faydalanılarak elde edilen tweet metninin polaritesi hesaplanmıştır. XML dosyası içindeki örnek satır aşağıdaki gibidir.

<word form="careful" wordnet_id="a-02456698" pos="JJ" sense="full of cares or anxiety" polarity="-0.5" subjectivity="1.0" intensity="1.0" confidence="0.9" />

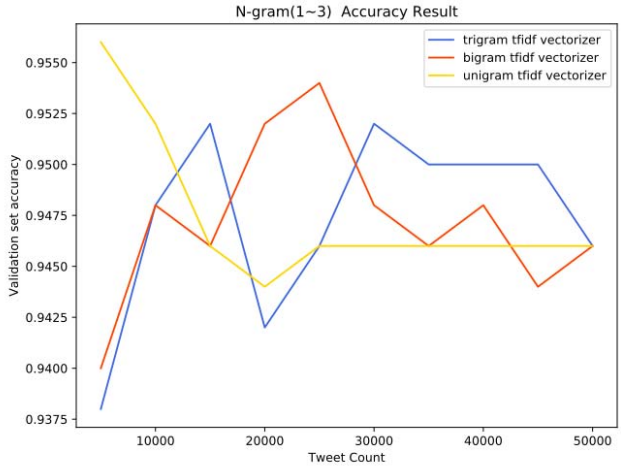
Hakkında tweet elde edilen beş para birimine ait polarite yüzdesi Tablo II'de gösterilmiştir.

TABLO II. Polarite Tablosu

Para Birimi	Olumlu	Olumsuz	Nötr
Bitcoin	%47,56	%45,47	%7,02
Ethereum	%48,36	%46,77	%4,87
Ripple	%41,68	%47,39	%10,93
Bitcoin Cash	%37,58	%53,08	%9,34
Litecoin	%41,36	%49,25	%9,39

Tablo II'de görüldüğü üzere, bu para birimlerinden en olumlu tweet alan %48,36 oranında Ethereum iken, en olumsuz yaklaşılan %53,8 ile Bitcoin Cash para birimi olduğu görülmektedir.

Tweetlerin polaritesi hesaplanırken, ayrıca n-gram (1,2,3) metodu ile metinler ayrılmış ve doğrulukları hesaplandığında ciddi bir fark olmadığı Şekil 2'te görülmektedir.



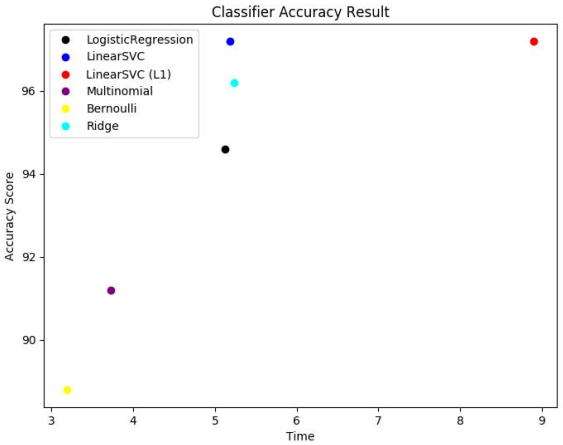
Şekil 2. N-Gram Doğruluk Skoru

D. Verilerin Sınıflandırılması ve Kullanılan Algoritmaların Karşılaştırılması

Metin sınıflandırırken Tablo III ve Şekil 3' te gösterildiği gibi birçok sınıflandırıcı kullanılmıştır. Kullanılan algoritmalar, doğruluk yüzdeleri, algoritmanın işlem süresi tablo III'de gösterilmiştir. Tabloya bakıldığında metin sınıflandırıcılardan Liner SVC ve Liner SVC (L1) 'in diğerlerine göre daha yüksek ve aynı doğruluk oranlarına sahip olduğu görülmektedir. Daha sonra sırasıyla Ridge, Logistic Regression, Multimodal, Bernoulli sınıflandırıcılarının %88 üzeri bir doğruluk oranı elde ettiği görülmektedir.

TABLO III. Sınıflandırıcı Doğruluk ve Süre Skorları

Sınıflandırıcı	Doğruluk	Süre(saniye)
Logistic Regression	%94,60	5,12
Linear SVC	%97,20	5,18
LinearSVC(L1)	%97,20	8,90
Multinomial	%91,40	3,73
Bernoulli	%88,80	3,19
Ridge	%96,20	5,23



Şekil 3. Sınıflandırıcı Doğruluk ve Süre Skorları

IV. SONUÇ

Bu çalışmada ilk defa günümüzdeki trend konulardan olan kripto para birimlerine karşı Twitter kullanıcılarının duyarlılık analizi yapılmıştır. Twitterdan elde edilen veriler anlamsız ve gereksiz verilerden arındırılmış ve polaritesi hesaplanmıştır. Literatürdeki bilindik sınıflandırıcılar kullanılarak yüksek doğruluk oranları elde edilmiştir. İleriki çalışmalarda yine güncel konular üzerinde lokasyon bazlı duyarlılık analizi yapılacaktır. Çalışmanın araştırmacılara fikir vermesi amaçlanmaktadır.

KAYNAKLAR

[1] Anonim, 2013. https://idc-cema.com/eng/events/50533-idc-big-data-and-business-analytics-forum-2013?g_clang=TR, Erişim Tarihi: 10 Aralık 2017.

[2] Mayer-Schonberger, V, Cukier, K, 2013. "Big Data: A Revolution That Will Transform How We Live, Work, and Think", Boston, New York

[3] Dirin, B. 2014. “Big Data Nedir?”, https://netvent.com/big-data-nedir/, Erişim Tarihi: 19 Aralık 2017.

[4] Laney, D, 2001. “3D Data Management: Controlling Data Volume, Velocity and Variety”. Tech. rep. Meta Group.

[5] Linch, M. 1998. https://en.wikipedia.org/wiki/Unstructured_data, Erişim Tarihi: 19 Aralık 2017.

[6] Bisk, 2017. “What is Big Data?”, https://www.villanovau.com/resources/bi/what-is-big-data, Erişim Tarihi: 20 Aralık 2017.

[7] Biehn, N. 2017. “The Missing V’s in Big Data: Viability and Value”. https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/, Erişim Tarihi: 21 Aralık 2017.

[8] Snijders, C., Matzat, U., Reips, 2012. "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science. 7: 1–5.

[9] Utkun, G., MICROSOFT Blog. 2012. “Büyük Veri Nedir”. http://blog.microsoft.com.tr/?p=4068. Erişim Tarihi 15 Aralık 2017.

[10] Baykal, A., 2006. “Veri Madenciliği Uygulama Alanları”, D.Ü. Ziya Gökalp Eğitim Fakültesi Dergisi 7, 95-107.

[11] Davenport, T, 2012. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/. . Erişim Tarihi: 10 Ocak 2018.

[12] James, J, 2017. “Data Never Sleeps”. https://www.domo.com/blog/data-never-sleeps-4-0/, Erişim Tarihi: 15 Kasım 2017.

[13] Hvistendahl, Mara, 2017. “Can ‘predictive policing’ prevent crime before it happens?”, http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens, Erişim Tarihi: 17 Aralık 2017.

[14] Domenico, T, 2013. “Clouds for Scalable Big Data Analytics”. In: IEEE Computer Society 46 , Sayfa: 98–101.

[15] Cavallo, A, 2017. “Scraped Data and Sticky Prices: Frequency, Hazards, and Synchronization”. Doktora Tezi. Harvard University.