T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK ANABİLİM DALI

BÜYÜK VERİ ARAÇLARI KULLANARAK SOSYAL MEDYADA HİS ANALİZİ YAPMA

YÜKSEK LİSANS TEZİ

HAZIRLAYAN: Mehmet Can ERDOĞAN
DANIŞMAN: Dr.Öğr. Üyesi Murat CANAYAZ

VAN-2019

T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK ANABİLİM DALI

BÜYÜK VERİ ARAÇLARI KULLANARAK SOSYAL MEDYADA HİS ANALİZİ YAPMA

YÜKSEK LİSANS TEZİ

HAZIRLAYAN: Mehmet Can ERDOĞAN

VAN-2019

KABUL VE ONAY SAYFASI

Elektrik-Elektronik Anabilim Dalı'nda Dr. Öğr. Üyesi Murat CANAYAZ danışmanlığında, Mehmet Can ERDOĞAN tarafından sunulan "**Büyük Veri Araçları Kullanarak Sosyal Medyada His Analizi Yapma**" isimli bu çalışma Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili hükümleri gereğince 07/01/ 2019 tarihinde aşağıdaki jüri tarafından oy birliği ile başarılı bulunmuş ve Yüksek Lisans tezi olarak kabul edilmiştir.

Başkan	: Prof.Dr. Naci GENÇ	İmza: .....
Üye	: Dr. Öğr. Üyesi Murat CANAYAZ	İmza: .....
Üye	: Dr. Öğr. Üyesi Murat DEMİR	İmza: .....
Üye	:.....	İmza: .....
Üye	:.....	İmza: .....

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 25.../01.../2019 tarih ve .....2019/6-I..... sayılı kararı ile onaylanmıştır.

...../...../20

Enstitü Müdürü

Prof.Dr. Suay GENÇ

Enstitü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Mehmet Can ERDOĞAN

ÖZET

BÜYÜK VERİ ARAÇLARI KULLANARAK SOSYAL MEDYADA HİS ANALİZİ YAPMA

ERDOĞAN, Mehmet Can
Yüksek Lisans Tezi, Elektrik-Elektronik Anabilim Dalı
Tez Danışmanı: Dr. Öğr. Üyesi Murat CANAYAZ
Ocak 2019, 67 sayfa

Sosyal medya araçlarının hayatımıza girmesi ile üretilen veri miktarı baş döndürücü boyutlara ulaşmıştır. Verileri analiz etmekte, geleneksel yöntemlerin artık yetersiz kaldığı 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, makine öğrenmesi, yapay zeka gibi kavramlar, daha fazla insanın ilgisini çekmektedir. Bu çalışmada, büyük veri araçları kullanılarak, en çok kullanılan kripto para birimleri ile makine öğrenmesi ve yapay zeka kavramlarına Twitter kullanıcılarının yaklaşımları ilk defa incelenmiştir. Twitter'dan elde edilen tweetler üzerinde anlamsız veriler temizlenmiş, kullanıcıların bu kavramlara olan yaklaşımları çeşitli sınıflandırıcılar kullanılarak analiz edilmiş ve sonuçlar gösterilmiştir.

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

ABSTRACT

SENTİMENT ANALYSIS IN SOCIAL MEDIA BY USING BIG DATA TOOLS

ERDOĞAN, Mehmet Can
Master Degree Thesis, Department of Electrical and Electronics Engineering
Supervisor: Assist. Prof. Dr. Murat CANAYAZ
January 2019, 67 pages

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, machine learning and artificial intelligence 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.

Key words: Big data, sentiment analyse, crypto currency, machine learning, text classification, social media, Twitter.

ÖN SÖZ

Bu tez çalışmasında verdiği desteklerden dolayı yardımlarını esirgemeyen danışmanım Van Yüzüncü Yıl Üniversitesi Mühendislik Fakültesi öğretim üyelerinden Dr. Öğr. Üyesi Murat CANAYAZ'a, ayrıca çalışmalarımda bana her zaman destek olan aileme ve değerli dostlarıma en içten teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELER LİSTESİ

Çizelge **Sayfa**

ŞEKİLLER LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklama
IoT	Internet of Things
TF-IDF	Term Frequency — Inverse Document Frequency
SVM	Support Vector Machine
IP	Internet Protocol
FTP	File transfer Protocol
XML	eXtensible Markup Language
NB	Naive Bayes
HTML	HyperText Markup Language
WWW	World Wide Web
SMS	Short message service
JSON	JavaScript Object Notation
NLP	Natural Language Processing
API	Application Programing Interface
IMAP	Internet Message Access Protocol
ML	Machine learning
DL	Deep Learning

1. GİRİŞ

İnsanoğlu, tarihi boyunca sürekli veri üretmiş ve sonraki nesiller için veya kendileri kullanmak üzere kayıt altına almıştır. İlk kayıtları, mağara duvarlarına çizilen şekiller olarak görecek olursak, başlarda çok az veri üretilmiş, zaman geçtikçe bu üretilen veri hızla artmıştır. Özellikle 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 cihazın internete erişim sağlayarak kendi aralarında veya daha büyük sistemlerle iletişimde olduğu bir ağdır. Mutfak gereçleri, güvenlik sistemleri, kameralar, fotoğraf makineleri, arabalar bu ağa bağlanmaya başlamakla birlikte ve önümüzdeki birkaç yıl içinde daha fazla cihazın bu listeye ekleneceği öngörülmektedir.

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 hakkında 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. Günümüzden 2020 yılına kadar, internetteki veri miktarının, her iki yılda yaklaşık iki katına çıkacağı düşünülmektedir. IDC, 2020 yılında, verilerin analiz edilmesi durumunda %33'ünün anlamlı bilgiler içereceğini tahmin etmektedir. İlk yıllarında, dijital evrenin yaklaşık yarısı sadece ABD ve Batı Avrupa'da iken gelişmekte olan pazarlar ise bu verinin %20'den azını oluşturmaktaydı, dijital evrenin gelişmekte olan pazarlardaki payı 2012 yılında %36'ya çıkmış ve 2020 yılına kadar ise %62'ye ulaşacağı tahmin edilmektedir (Anonim, 2013).

Yukarıda belirtilen veri boyutu birimlerini aşağıdaki (Çizelge 1.1) incelendiğinde daha iyi anlaşılacaktır.

Çizelge 1.1. Veri kapasite birimleri (Anonim, 2017a)

İşlemci veya Sanal Depolama	Disk Depolama
8 Bit = 1 Byte	8 Bit = 1 Byte
1024 Byte = 1 Kilobyte	1000 Byte = 1 Kilobyte
1024 Kilobyte = 1 Megabyte	1000 Kilobyte = 1 Megabyte
1024 Megabyte = 1 Gigabyte	1000 Megabyte = 1 Gigabyte
1024 Gigabyte = 1 Terabyte	1000 Gigabyte = 1 Terabyte
1024 Terabyte = 1 Petabyte	1000 Terabyte = 1 Petabyte
1024 Petabyte= 1 Exabyte	1000 Petabyte= 1 Exabyte
1024 Exabyte = 1 Zettabyte	1000 Exabyte = 1 Zettabyte
1024 Zettabyte = 1 Yottabyte	1000 Zettabyte = 1 Yottabyte
1024 Yottabyte = 1 Brontobyte	1000 Yottabyte = 1 Brontobyte
1024 Brontobyte = 1 Geopbyte	1000 Brontobyte = 1 Geopbyte

1000, 1024 ile değiştirilebilir ve yine kabul edilebilmektedir. Söz konusu saklama türüne bağlı olarak bu standartların her ikisinin de doğruluğu kabul görmüştür(Anonim, 2017a).

1.1 Büyük Veri

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 ve bir düzende tutulması ilk olarak astronomi ve genetik alanında olmuştur. Büyük veri kavramı da ilk olarak bu alanlarda kullanılmış, ancak daha sonra birçok alanda verilerin artması ile bu kavram daha sık kullanılmaya başlanmıştır. Büyük veri artık hayatımızın neredeyse her alanında kendini göstermeye başlamıştır (Mayer-Schonberger, 2013).

Veriler bu kadar artarken, bütün veriler anlamlı veya yorumlanabilir durumda değildir. Bu verileri analiz etmek ve 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 bir olgudur. Bizler de bu olguya destek olmakta ve "Büyük Veri" olarak adlandırdığımız bu ortama sürekli veri akışının sağlanmasında katkı sağlamaktayız (Dirin, 2017).

2000'li yıllarda büyük veri üç bileşenli olarak tanımlanmış ve İngilizce baş harflerinden dolayı 3V (Volume-Hacim, Velocity-Hız, Variety-Çeşitlilik) olarak isimlendirilmiştir (Laney, 2001).

Volume (Hacim veya veri büyüklüğü). Üretilen verinin miktarı ve saklanan verinin boyutunu temsil eder. Üretilen veri miktarına göre o verinin aslında büyük veri olup olmadığına karar verilir.

- Veriyi birçok farklı kaynak üretmektedir. IDC (International Data Corporation) araştırmalarına göre. veri 44 farklı cihazdan üretilmektedir. Algılayıcılar, süper bilgisayarlar, kişisel bilgisayarlar, sunucular, arabalar, uçaklar, v.b.
- IDC istatistiklerine göre 2013'ten 2020'ye veri miktarı 4.4 trilyon gigabayttan 44 trilyona çıkacaktır.
- IDC istatistiklerine göre 2013'te dijital evrenin %20'si internet ortamında işlem görürken 2020'de bu oran %40 olacaktır.

Velocity (Hız). Büyük veri, yukarı doğru ivmeli bir hızla üretilmekte ve bu veriler çok kısa sürede devasa boyutlara ulaşmaktadır. Hızla büyüyen veri, o veriyi kullanan işlem sayısını ve çeşitliliğini de aynı hızda artmaktadır. Hem yazılım hem de donanım olarak bu yoğunluğa cevap verilebilmektedir.

Variety (Çeşitlilik). Üretilen veriler genel olarak belli bir düzende ve birçok farklı cihazdan elde edilen veri biçimlerinden oluştukları için, verilerin sistemli ve birbirlerine dönüştürülebiliyor olmaları gerekmektedir.

- Dijital evrenin %70-%80 düzenli olmayan veriler oluşturur.

- Anlamlı bilgilerin %80-%90 düzenli olmayan verilerden elde edilir (Linch,1992).
- 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 (Bisk ve Biehn, 2017).

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

Value (Değer). 5V'nin en önemli aşaması sayılabilecek "Value" katmanı, diğer katmanları geçip büyük veri nitelemesini kazanan verilerin ne kadar değerli olduğunun incelendiği katmandır.

1.2. Veri Madenciliği

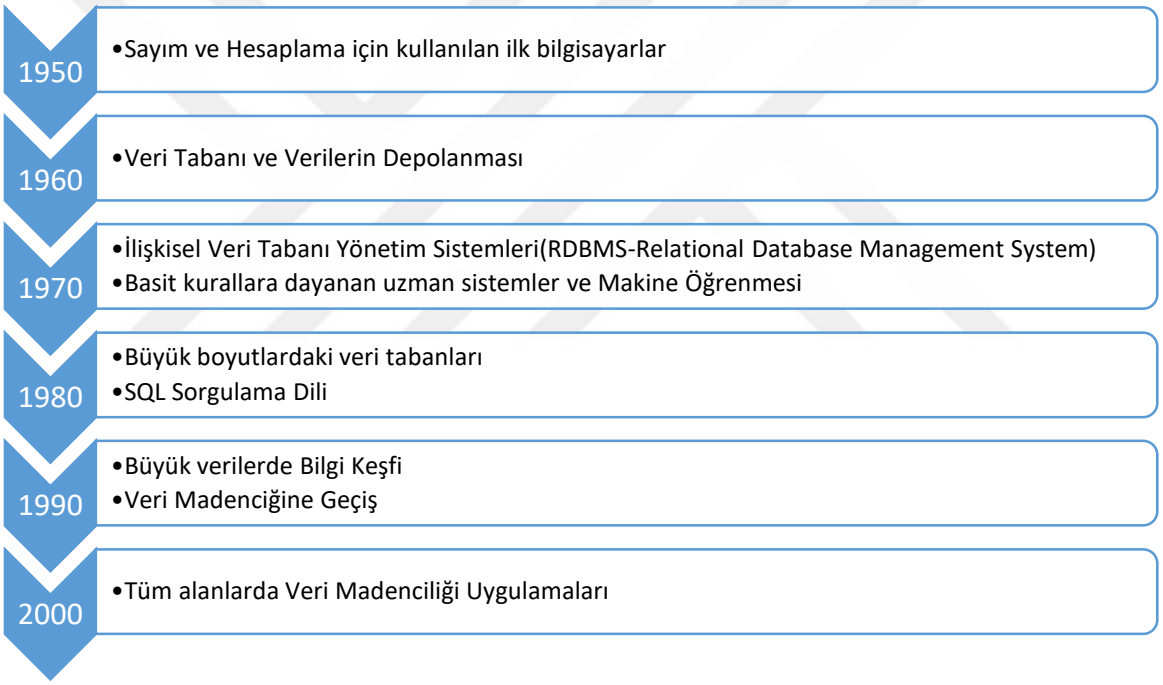
Veri madenciliği, günümüzde devasa boyutlara ulaşan verilerden özet, anlamlı ve işe yarar bilginin elde edilmesidir. Bu bilgiyi elde edebilmek için veriler üzerinde sınıflandırma, veri özetleme, veriler üzerinde yapılan değişikliklerin analizi ve verilerde yapılan hataların tespiti gibi işlemler yapmak gerekir (Savaş ve ark., 2012).

İlişkisel veri tabanı yönetim sistemleri günümüzde hem büyük boyutlara ulaşan hem de veri türünde çok fazla çeşitlilik gösteren verileri analiz etmede çoğu zaman yetersiz kalmaktadır.

1950'li yıllarda bilgisayarlar sayım ve temel matematiksel işlemler için kullanılmaya başlamıştır. 1960'lı yıllarda basit veri tabanı sistemlerine ilk veri kayıtları yapılmakla birlikte 1970'li yıllarda ilişkisel veri tabanı yönetim sistemleri (RDBMS-Relational Data Base Management Systems) hayatımıza girerek basit düzeydeki kurallara dayanan uzman sistemler ile ilk makine öğrenmesi işlemleri gerçekleştirilmiştir. 1980'li yıllarda veri tabanlarının boyutları artmış, veri ambarı terimi kullanılmaya başlanmıştır. Veri üzerinde daha hızlı ve kolay bir şekilde işlem yapabilmek için yapısal sorgulama dili

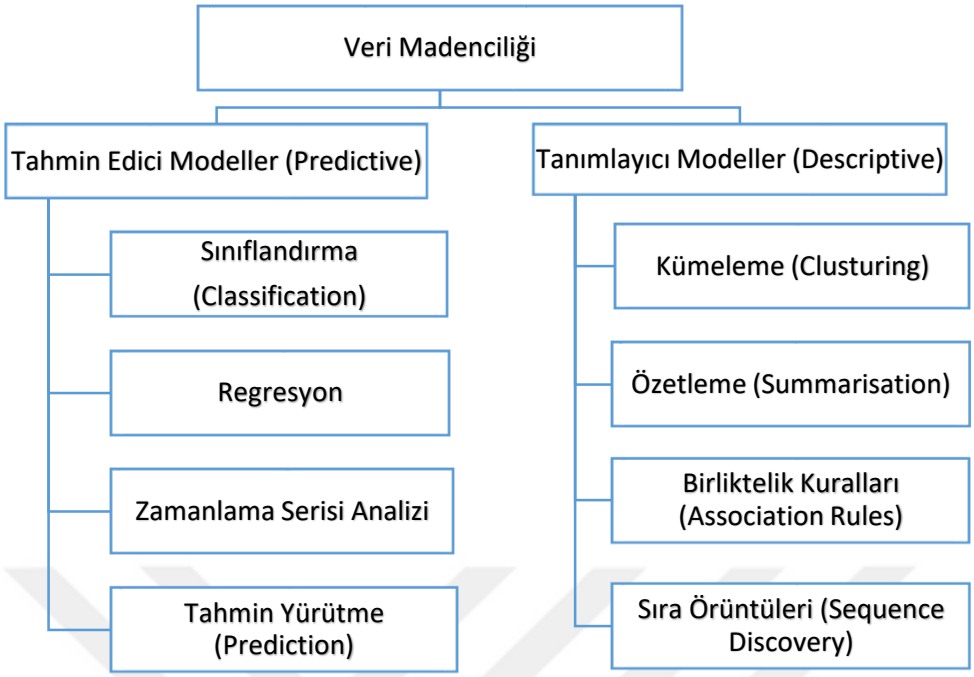
(SQL- Structured Query Language) geliştirilmiştir. 1990'lı yıllarda, ilk veri madenciliği işlemleri yapılmış, bu veri ambarları üzerinde bilgi keşfi dönemi başlamıştır. 2000'li yıllarda artık sadece veri tabanları üzerinde değil birçok disiplin (bankacılık ve finans, pazarlama, mühendislik çalışmaları, sigortacılık, sağlık, biyomedikal ve DNA, imalat, internet, telekomünikasyon, eğitim vb.) ile de ilişkili bir şekilde veri madenciliği dönemi başlamıştır. İstatistik, veri tabanı teknolojisi, makine öğrenmesi ve veriyi görselleştirme (Visualisation) işlemleri veri madenciliğinin alt başlıkları olarak görülebilir (Aydın,2007).

Şekil 1.1'de gösterilen veri madenciliğinin tarihsel gelişimi de bu fikri desteklemektedir.



Şekil 1.1. Verinin tarihsel gelişimi.

Veri madenciliği, veriler üzerinde temelde iki işlemi yerine getirmektedir (. Şekil 1.2). Bunlardan ilki, sadece eldeki verileri ele alarak özet bir bilgi çıkarmaya çalışan "Tanımlayıcı (Descriptive) Modeller" ve ikincisi, eldeki verilerden yol çıkarak, birçok farklı algoritma ile gelecek verilerin nasıl olacağının en iyi tahminini yapmaya çalışan "Tahmin Edici (Predictive) Modeller"dir (Aydın,2007).



Şekil 1.2. Veri madenciliği modelleri.

Tahmin Edici (Predictive) Modeller, sonucu bilinen unsurlar veya analizi yapılmış ve sonuca ulaştırılmış verileri temel alarak, sonucu bilinmeyen unsurlar ile daha yeni işlenecek veriler üzerinde bir tahmin yürütmeye çalışmaktadır. Bu tahminin doğruluğunu algoritmalar ile en iyi dereceye çekmeye çalışılmaktadır.

Tanımlayıcı (Desciptive) Modeller, veriler arasındaki örüntüyü ve ilişkiyi tanımlamaya çalışmaktadır. Verilerin özniteliklerini ve bir biri ile olan bağıntılarını inceleyip bir tanım oluşturarak özet bilgiye ulaşmaya gayret göstermektedir.

1.3. Makine Öğrenmesi

Makine öğrenimi, bilgisayarların, veri üreten algılayıcılardan (sensör) veya veri tabanlarından elde ettiği veriden yola çıkarak öğrenimini olanaklı kılan algoritmaların tasarım ve geliştirme süreçlerine dayalı bir terimdir. Makine öğreniminin amacı bilgisayarlara, veriler arasındaki bağlantıları algılama ve veriye dayalı mantıklı kararlar verebilme yeteneği kazandırmaktır. Bundan dolayı, makine öğrenimi aynı zamanda istatistik, veri madenciliği, yapay zekâ gibi alanlar ile yakından ilişkilidir (Anonim, 2017b).

Makine öğrenimi, verilerdeki değişimleri öğrenebilen ve bunlara uyum sağlayabilen yapay zekanın bir parçasıdır. Bu nedenle olası her durum için bir algoritma kullanmaya gerek kalmaz. Geçmiş deneyimlere bakılarak yeni duruma uyum sağlamaya çalışılır. Bir bilgisayar için geçmiş deneyim veri tabanlarıdır. Dolayısı ile makine öğreniminin bir veri kümesine ihtiyacı vardır.

Makine öğreniminin gözetimli ve gözetimsiz olmak üzere iki türü vardır. Gözetimli öğrenme, daha önce etiketlenmiş gözlemlere (analiz edilmiş verilerden, örneğin tweet olumlu mu olumsuz mu veya mail spam mi değil mi gibi) dayanan bir öğrenme sürecidir. Örneğin tweette olumsuz olarak etiketlenen kelimeler geçiyorsa o tweet olumsuzdur denmektedir. Bu sayede tweetler üzerinde sınıflandırma (olumlu/olumsuz/nötr) yaparak bir sonuç tahmini (regresyon) yapmaya çalışılır. Örneğin. "Bitcoin yükselişte iken hakkında atılan tweetler olumlu olmalıdır". Gözetimsiz öğrenme ise daha sınıflandırma yapılmamış verileri esas alan öğrenme sürecidir. Burada algoritmanın veriler arasındaki örüntüyü bulması beklenir. Bulunan örüntüye göre öznitelik olarak birbirine yakın verilerin aynı grupta yer alması sağlanmaktadır.

1.4. Web Kavramı

1960'lı yılların başında ilk hali ortaya çıkan internet, başlarda sadece askeri anlamda kullanılmış olan, ancak daha sonra birçok bilgisayarın ve günümüzde birçok cihazın(telefon, araba, ev aletleri...) bağlı olduğu bir uluslararası ağdır. İnternetin sadece askeri kullanımın dışına çıkmasıyla birlikte, birçok yeniliği ve kolaylığı beraberinde getirdiği bilinmektedir. Herkesin ulaşacağı sunucu bilgisayarlarda bulunan ve estetik olarak daha iyi bir görünüme sahip olan Web sayfaları ile kullanım alanı ve erişimi ciddi bir ivme kazanmıştır.

Web (World Wide Web) kavramı ortaya çıkmadan önce sadece yazılı metinler paylaşılmaktaydı. Buna rağmen o dönemin koşullarında, bilgi paylaşımı konusunda çok önemli bir yere sahip olduğu görülmektedir. Web'in ortaya çıkmasıyla görsellik ve kullanım açısında daha yetenekli paylaşım ortamları ortaya çıkmıştır. İnterneti sıradanlıktan kurtarmakla kalmayıp metin dışında üretilen fotoğraf, video ve ses dosyalarının paylaşımına da olanak sağlamıştır (Peter, 2010).

Web'in ilk hali, Berners-Lee'ye göre "salt okunur web" olan web 1,0'ı temsil etmektedir. Başka bir deyişle, ilk web, bilgi aramamıza ve okumamıza izin verirken kullanıcı etkileşimi veya içerik oluşturma konusunda web kullanıcılarına bu imkanı sağlamaz iken, zaman ve mekan sınırlaması olmadan dünyanın herhangi bir yerinden bilgiye ulaşmak büyük ve önemli bir durum olarak görülmektedir (Getting,2017).

Web 1.0, sadece web sayfalarında var olan bilgilere erişmek, sayfa sahipleri tarafından sağlanan içerikleri okumak, program ve dosyaları indirmek için kullanmaktaydı (Alikılıç, 2011).

1990'ların sonuna doğru, Web sayfalarında etkileşime olanak sağlayan birçok uygulama hayata geçmiş, önemli sosyal etkileşim ve oluşumların zemini haline gelmiştir. 1999 yılında DaveWinner'ın kullanıcıların ziyaret ettikleri sitelerde içerik oluşturmasına (yorum yapmasına) imkân veren yazılımı, Web'i salt okunur olmaktan çıkarmıştır. Web, Tim Berners-Lee'nin başından beri öngördüğü 'Okunabilir/Yazılabilir Web'e (Read/Write Web) dönüşmüştür (Karabulut, 2009).

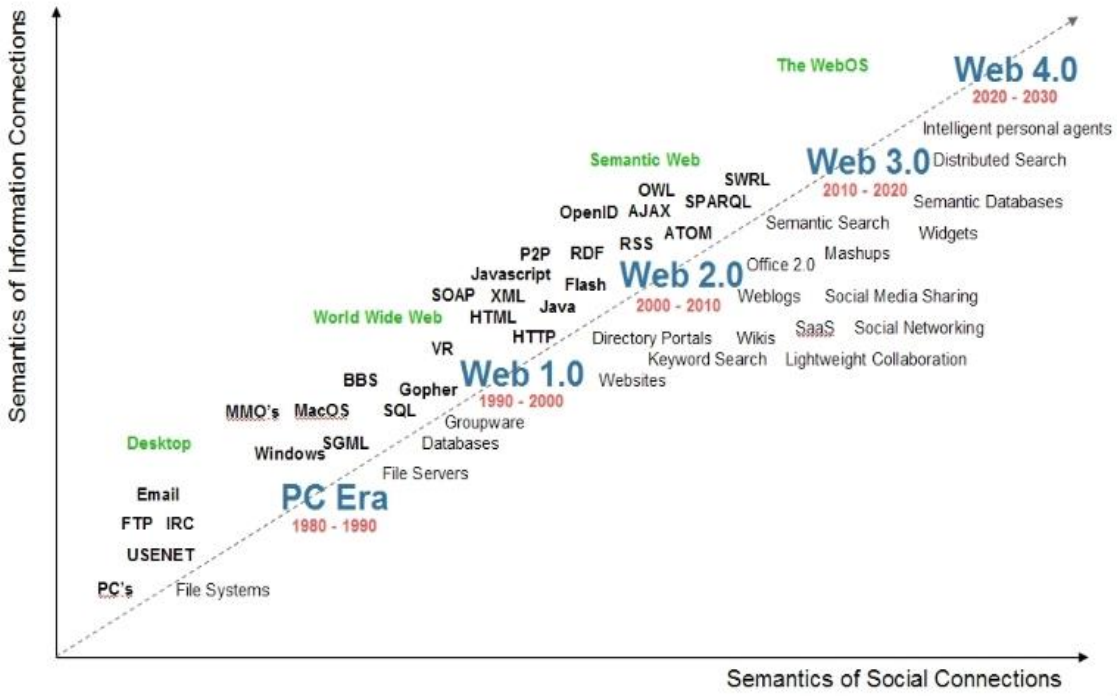
Günümüzde, Web 2.0'ın ya da Berners-Lee'nin tarif ettiği "okuma-yazma" webinin ilk hallerini görmekteyiz. Web kullanıcıları artık web içeriğine katkıda bulunmakta ve diğer web kullanıcılarıyla etkileşim kurabilmektedir. Web sayfaları tek yönlü olmaktan çıkıp çift önlü etkileşime de imkan sağlamaktadır. Örnek olarak, kullanıcı gönderimlerine dayanan YouTube ve MySpace gibi paylaşım sayfaları, bilgiye katılmak isteyen veya ürettiği bilgiyi paylaşmak isteyen web kullanıcılarının bu isteklerine yanıt olabilmektedir.

Günümüzde web sayfalarının büyük çoğunluğu Web 2.0'ı kullanmaktadır. Ancak internetteki gelişim sayesinde Web 3.0'a geçiş kaçınılmaz olacaktır. Şu anda bazı web sayfaları veya uygulamalar Web 3.0'a geçiş yapsa da tamamen geçiş yaptığımızı söylemek mümkün değildir.

Tim Berners-Lee'nin "okuma-yazma-yürütme" olarak tanımladığı Web 3.0, soyut olarak tanımlamanın zor olduğu görülmektedir. Semantik Web (anlamsal-anlambilimsel) olarak tanımlanmakta ve kullanıcıların da içerik üretimine katılmasıyla her yıl artan bilgilerin birbiri ile doğru ilişkilendirerek kullanıcın önüne koyacak sistemler geliştirmektedir. Anlamsal web kısaca, arama motorları gibi uygulamaların, web'de farklı kaynaklarda var olan bilgiyi anlamlı bir şekilde okuyup değerlendirebileceği, aradaki

ilişkiyi görüp bir araya getirebileceği aşamayı ifade etmektedir. Bilgisayar temelli bir yapıya sahip olmasından bundan dolayı Semantik Web için aynı zamanda "Web of Data" kavramı da kullanılmaktadır. Her ne kadar Semantik Web uygulamalarının çoğu henüz geliştirme aşamasında olsa bile bazı uygulamalarda şimdiden hayata geçirilmiştir (Ege, 2011).

Web'in sürümleri ve hangi sürümde hangi teknolojilerin kullanıldığı bilgisi aşağıdaki şekilde (Şekil 1.3) gösterilmiştir.



Şekil 1.3. Web'in gelişimi (Pileggi ve ark., 2012).

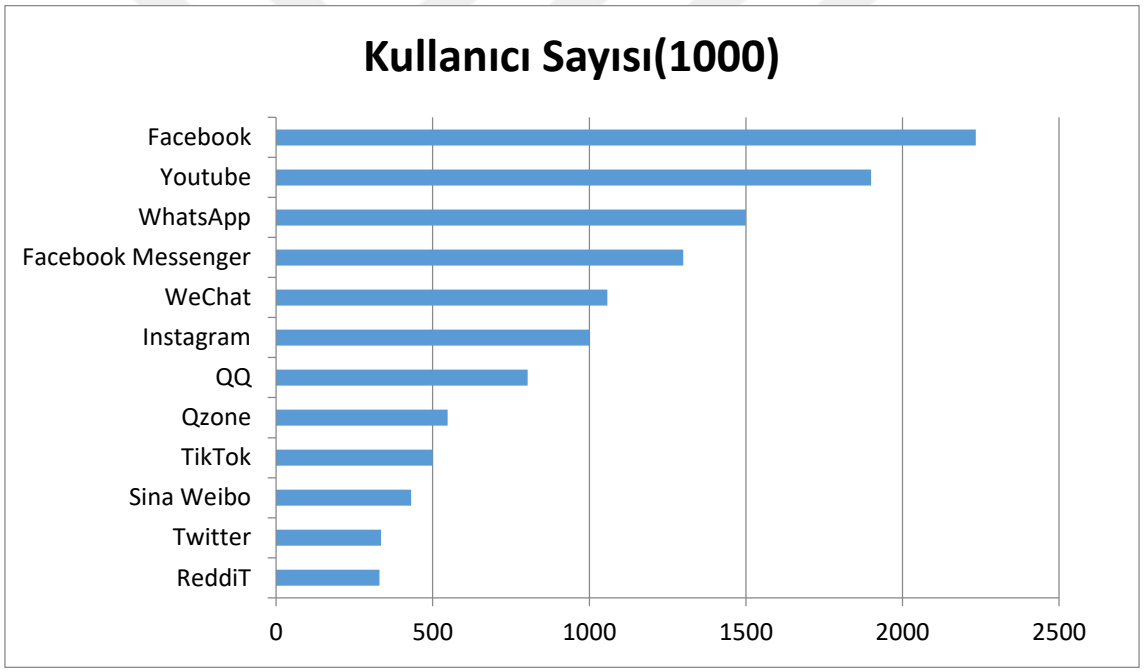
1.5. Sosyal Medya

Sosyal medya, kullanıcılar tarafından üretilen bilginin anlık ve çift yönlü paylaşımı demektir (Çelik, 2018). İlk olarak 1960 yılında temelleri atılan internet, geçen zaman içinde bilgi teknolojilerinin de gelişmesiyle, insanlar arasında bireysel iletişim ihtiyacı doğurmuştur. 1970'li yıllarda Ward Christensen adındaki yazılımcı kendi

arkadaşları ile iletişim kurmak için CBBS (Computerized Bulletin Board System-Bilgisayarlı Bülten Tahtası Sistemi) ismini verdiği bir program yazmıştır. Bu program, kendisi ve diğer bilgisayar meraklılarının birbirileri arasında bilgi alışverişinde bulunmalarına olanak sağlamıştır (Scott, 2005).

Güncel anlamdaki sosyal medya ise Harvard Üniversite öğrencisi Mark Zuckerberg adında bir öğrencinin, okul öğrencilerini bir araya getirip iletişim kurmalarını sağlayan Facebook'u kurmasıyla başlamıştır. Başlarda Harward Üniversitesi özelinde mezunları bir araya getirmeyi amaçlayan Facebook, daha sonra bütün dünya okullarına yayılmış ve günümüzde iki milyarın üzerinde kullanıcıya sahiptir.

Kullanıcı sayısında göre statista.com adresinden alınan ve 2018 yılı itibariyle aktif kullanıcı sayısını gösteren değerler (Şekil 1.4) aşağıdaki gibidir.



Şekil 1.4. Sosyal medya kullanıcı sayıları (Anonim, 2018a).

İnternetin yaygınlaşması ve erişim sağlayan kişilerdeki artış sebebiyle birçok sosyal medya platformu günümüzde milyonlarca kullanıcıya, kullanıcılar da sosyal medya araçlarına erişim sağlamaktadır.

Günümüzde insanların neredeyse üçte ikisi sosyal medya araçlarından en az birini kullanmaktadır. Bu kadar kitlenin olduğu bir ortamda gelişmeler çok hızlı olmakta ve hem sosyal medya araçları insanların alışkanlıklarına yön vermekte hem de insanlar

sosyal medyanın yeteneklerine katkıda bulunmaktadır. Birçok hizmet sunmasına rağmen sosyal medya siteleri belli konulara yoğunluk vererek yeni kategorilerin oluşmasında rol oynamaktadır.

En çok kullanılan sosyal medya kategorileri ve insanlar tarafından en çok kullanılan siteler aşağıdaki gibi listelenebilir (İnanç, 2018).

- Bloglar. Blogger, LiveJournal, Open Diary, TypePad, WordPress.
- Mikro Bloglar. FMyLife, Jaiku, Plurk, Twitter, Tumblr, Posterous, Yammer.
- Sosyal Ağlar. Bebo, BigTent, Elgg, Facebook, Geni, Hi5, LinkedIn, MySpace.
- Sosyal Haber. Digg, Mixx, Reddit.
- Fotoğraf Paylaşımı. deviantArt, Flickr, Photobucket, Picasa.
- Video Paylaşım. YouTube, Viddler, Vimeo, Sevenload, Dailymotion.

Blog kavramı, İngilizce "Web" ve "Log" kelimelerinin kaynaşması ile oluşmuş, web günlüğü anlamında kullanılmaktadır. Bunu kullanabilmek için ciddi teknik bilgiye ihtiyaç duyulmadan, dosya aktarımı ve özel yazılım bilgisi gerektirmeden web sayfası oluşturmayı ve paylaşmayı sağlayan bir sistemdir. Teknik bilginin yanında yüksek maliyet de gerektirmeyen, sadece bilgisayar kullanmasını bilenlerin oluşturabileceği sistemlerdir (Karcıoğlu, Kurt, 2009).

Bloglar, genellikle, son yazılan yazılar başa geldiğinden, yeniden eskiye konuları yorumlayan yazı dizilerinden oluşmaktadır. Yazının başlığı, içerik ve sonunda yazar bilgileri ve yazının yayınlanma tarihi bulunmaktadır. Yayıncının tercihine göre, diğer kullanıcılar yazının altına yorum yazabilmektedir. Yazıları yazanlar için artık "yayıncı"dan çok "blogger" terimi kullanılmaktadır (Alper, 2012).

Bloglar, uzun yazı dizilerinden oluştuğundan, anlık gündemi takip etmekten uzak kalmıştır. Mikroblog siteleri ise her yazı (blog) için karakter sınırlaması getirerek kullanıcıları anlık veya günlük olaylar hakkında kısa fikirlerini paylaşan daha dinamik siteler olmuştur.

1.6. Twitter

2006 yılında Jack Dorsey'in arkadaşları ile beraber kurduğu Twitter, kendi anlık durumlarını, bulundukları yeri ve yaptıkları aktiviteleriyle ilgili güncelleme yapmayı sağlayan, kentli yaşam tarzına ait bir araç olarak geliştirilmiştir (Demirbas ve Akcora, 2010).

Twitter, hem "kuş sesleri" hem de "yararsız bilgilerden oluşan kısa patlama anlamına gelmektedir (Sarno, 2009). Twitter markasının yüzü olan kuş, atılan tweetlerin yararsız kuş cıvıldamalarına işaret etmekteydi. Ancak Jack Dorsey'e göre, size anlamsız gelen cıvıltı, başka kuşlarca anlaşılmaktadır.

İlk planlamada, sadece kendi ekip arkadaşlarına durumu ile ilgili bilgi içeren SMS olarak planlanan Twitter, yazılan tek SMS'in bütün grup elemanları tarafından okunması amaçlanmıştır. 160 karakter olan SMS boyutunun 20 karakteri gönderen kişi için ayrılmış, kalan 140 karakter ise durum paylaşımı için planlanmıştır. Twitter'ın 140 karakter olması buradan gelmektedir. Bu yüzden bazı kaynaklarda Twitter için "internetin SMS'i" benzetmesi yapılmaktadır.

İlk başlarda sadece iş arkadaşlarına nerede ve ne yaptığı ile ilgili bilgi verme fikri ile başlatılan ve sadece SMS olarak düşünülen Twitter, Jack Dorsey'e göre, "Nerede olursam olayım, telefonumu elime alsam ve ne olup bittiğini paylaşsam ve aynı anda herkesten de haber alsam nasıl olur?" sorusuna cevap vermeye çalışmıştır.

2017 ilk çeyreği itibariyle 328 milyon aktif kullanıcısı bulunan Twitter, aynı yılın sonlarına doğru (26 Eylül 2017) bazı diller hariç (Japonca, Çince ve Korece) diğer bütün dillerde 140 olan karakter sayısını 280 olarak değiştirmiştir (Kwak ve ark., 2010).

2006 yılından günümüze çok ciddi gelişme gösteren Twitter, milyonlara ulaşan kullanıcılarının 140 karakterlik bildirimlerinin ya da "tweet"lerinin olduğu bir mikroblog platformudur. Veri tabanının açık olması ve birçok kaynaktan kolayca erişilebilirliği sayesinde birçok çalışma için veri kaynağı olmuştur (Poynter, 2012).

Kullanıcı sayısının fazlalığı, atılan tweetlerin sıklığı ve her bir tweet'in tuttuğu bilgiler göz önüne alındığında, Twitter, birçok alanda iyi bir veri kaynağı olabilmektedir. Yazılan bir tweet'e ait birçok bilgi tutulmaktadır. Duyarlılık analizinde kullanılmak üzere en çok kullanılan veya kullanılabilecek tweet nesnesinin özellikleri Çizelge 1.2'de gösterilmiştir.

Çizelge 1.2. En çok kullanılan tweet bilgileri

Özellik	Veri Tipi	Açıklama
created_at	String	Hangi tarihte oluşturulduğu bilgisi
Id	Int	Her tweetin benzersiz kodu
Text	String	Tweet metninin kendisi
Source	String	Hangi cihazdan yazıldığı
User	Object	Şekil 1.5'te ayrıntıları gösterilmiştir.
coordinates	Coordinates	Tweet'in yazıldığı koordinat
Place	Place	Tweet'in yazıldığı coğrafi yer

Tweeti yazan kişiye ait id, isim, bulunduğu, konum gibi birçok bilgi tutulmaktadır. Kullanıcı nesnesinin JSON formatındaki gösterimi Şekil 1.5'teki gibidir.


```
{
    "user": {
        "id": 6253282
        "id_str": "6253282"
        "name": "Twitter API"
        "screen_name": "twitterapi"
        "location": "San Francisco, CA"
        "url": "https://dev.twitter.com"
        "description": "The Real Twitter API."
        "protected": true
        "verified": false
        "followers_count": 21
        "friends_count": 32
        "listed_count": 9274
        "favourites_count": 13
        "statuses_count": 42
        "created_at": "Mon Nov 29 21:18:15 +0000 2010"
        "geo_enabled": true
        "lang": "en"
        "lang": "msa"
        "lang": "zh-cn"

    }
}
```


Şekil 1.5. Kullanıcı nesnesi (user object).

1.7. Kripto Para

Kripto (crypto) ve Para (currency) kelimelerinin bir araya getirilmesiyle yeni türetilen bu kavram şifreli para anlamına gelmektedir. Kripto para; gerçek hayatta kullanılmayan sadece internet ortamında olan, herhangi bir merkezi otoriteye veya banka gibi aracı kurumlara bağlı olmayan sanal para birimi olarak ifade edilmektedir.

Bitcoin'in artan cazibesi ve ticari kabulü nedeniyle, değer oluşumunu etkileyen faktörleri anlamaya çalışmak gittikçe önem kazanmıştır. Öncelikle, var olan tüm bitcoinlerin değeri yaklaşık 7 milyar dolar ve 60 milyon dolar değerindeki bitcoine ait el değişimi gerçekleşmektedir. Son birkaç yılda, hızla büyüyen Bitcoin, ticari amaçlı olarak iyi bir pazar payı edinerek, dijital para birimlerinin yükselen varlık sınıfı olarak görülmektedir. Bitcoin ve diğer dijital para birimleri sadece listelenmiş ve tezgah üstü bir pazar değil, aynı zamanda yeni bir türev piyasadır. Bu nedenle, Bitcoin ve ilgili kripto para değerlerine değer verme yeteneği, meşru bir finansal varlık olarak kurulmasında kritik hale gelmektedir. Bu konu sadece finans ve ekonomi alanlarında değil, aynı zamanda bilgisayar bilimleri, bilgi sistemleri ve uygulamalı kriptografi ile kesişmektedir (Hayes, 2017).

Bitcoin veya diğer dijital para birimi platformlarının geleceği, geleneksel veya eski finansal sistemlere potansiyel olarak zarar verebilmektedir. Parasal işlemleri bankacılık sektöründen ve parasal otoriteden etkili bir şekilde uzaklaştırır ve bunun yerine sadece teknolojiyi kullanarak güven kurmaktadır. Bankalar arası ve finanse edilmemiş finansal hizmetlere erişim sağlayabilir, son derece düşük maliyetli para transferleri ve devlet sınırlarını aşan hizmetlere izin verebilmektedir (Hayes, 2017).

Kripto para kavramıyla beraber ismi daha çok duyulmaya başlanan Blockchain, kısaca şifrelenmiş işlemlerin takibini sağlamak amacıyla tutulan dağıtık veri tabanı olarak tanımlanmaktadır. Zincirleme bir modelle inşa edilen, takip edilebilen ama kırılamayan Blockchain teknolojisi, bir merkeze veya otoriteye bağlı olmaksızın işlem yapmaya izin vermektedir. Böylece banka gibi aracı kurumları aradan çıkararak kişiden kişiye işlemler sağlanabilmektedir. Günümüzde var olan, kişiden kişiye veya noktadan noktaya, müzik, video veya dosya paylaşım programlarına benzetilebilmektedir.

Bitcoin ve blockchain teknolojisinin ekonomik, politik, insani ve yasal sistem faydaları, bunun toplumun ve faaliyetlerinin tüm yönlerini yeniden yapılandırma kapasitesine sahip olabilecek son derece çarpıcı bir teknoloji olduğunu açıkça ortaya koymaktadır. Blockchain farklı türleri olan üç kategoriye ayrılmaktadır. Blockchain 1.0, 2.0 ve 3.0. Blockchain 1.0 para birimi, para transferi, havale ve dijital ödeme sistemleri gibi nakit ile ilgili uygulamalarda kripto para birimlerinin dağıtımıdır. Blockchain 2.0, basit nakit işlemlerden daha geniş olan blockchain'i kullanan ekonomik, piyasa ve mali uygulamaların tümünün kontratıdır. hisse senetleri, tahviller, vadeli işlemler, krediler, ipotekler, başlıklar, akıllı mülkler ve akıllı sözleşmeler. Blockchain 3.0, özellikle devlet, sağlık, bilim, okur yazarlık, kültür ve sanat alanlarında olmak üzere para, finans ve piyasaların ötesindeki blockchain uygulamalarıdır (Swan, 2015).

1.8. Duyarlılık Analizi

Duyarlılık veya duygu analizi, belirli bir konu hakkındaki yazılı veya sözlü fikirlerin anlaşılması için yapılan işlemlere denir.

Fikir madenciliği olarak da tanımlanabilen duyarlılık analizi, metnin içinde görüş belirlemeye ve çıkarmaya çalışan "Doğal Dil İşleme" (NLP- Natural Language Processing) içindeki bir alandır. Bu süreç temelde üç nitelik üzerine eğilmektedir (Anonim, 2017c);

Polarite. Görüşün olumlu veya olumsuz olma durumunu yani metnin pozitif veya negatif olma durumunu gösterir.

Konu. Hakkında fikir çıkarılacak konuyu gösterir.

Görüş Sahibi. Konu hakkında görüşlerini ifade eden kişileri veya kuruluşları ifade eder.

Analizi yapılacak metin temelde iki kısma ayrılabilir. gerçekler ve görüşler. Gerçekler bir konu hakkındaki *nesnel* ifadelerdir. Görüşler ise genellikle insanların duygularını, değerlendirmelerini ve bir konuya yaklaşım durumlarını açıklayan *öznel* ifadelerdir. Dolayısı ile bir metnin duyarlılık analizi yapılırken öncelikle metnin öznel-

nesnel ayırımının yapılması gerekmektedir. Örneğin "güneş doğudan doğar" cümlesi nesnel bir metin olup polarite işlemine tabii tutulmaması gerekmektedir. Fakat "sarı rengi çok güzeldir" cümlesi öznel ve bir konu hakkında olumlu görüş bildiren bir cümledir.

Duygu analizi, görüş sahibinin belirli bir konu hakkındaki fikrinin olumlu veya olumsuzluğuna bakar. "Niyet Analizi" ise görüş sahiplerinin söylediklerinden ziyade görüşleri ile ne yapmak istediklerini temel olarak almaktadır. Daha anlaşılır olması için güncel konulardan birkaç örnek vermek gerekirse;

"Müşteri desteğiniz çok kötü. Telefonda 20 dakika bekletildim ".

"Bu yazıcının nasıl çalıştığını bilmek istiyorum".

"Bu işlemi yapmama yardım edebilir misin?"

Birinci cümlede bir şikayet, ikinci cümlede bir soru ve üçüncü cümlede bir talep olduğu görülebilir.

Bilgi toplama davranışımızın önemli bir parçası, diğer insanların ne düşündüğünü bulmak olmuştur. Çevrimiçi gözden geçirme siteleri ve kişisel bloglar gibi kanaat bakımından zengin kaynakların kullanılabilirliği ve popülaritesinin artmasıyla, insanlar artık başkalarının fikirlerini araştırmak ve anlamak için bilgi teknolojilerini aktif olarak kullanabilecekleri yeni fırsatlar araştırmaktadır. Görüş, düşünce ve öznelliğin metin içinde hesaplanmasıyla ilgilenen fikir madenciliği ve duyarlılık analizi alanındaki ani patlama olayları (anlık ilgi), en azından kısmen yeni sistemlere olan ilginin artmasına doğrudan bir tepki olarak ortaya çıkmıştır (Pang 2008).

"Diğer insanlar ne düşünüyor?" cümlesi, karar verme sürecinde hepimiz için önemli bir bilgi parçası olmuştur. Web ve sosyal medya kullanımından önce, sadece etrafımızdaki bir kaç kişiden bir konu hakkında görüşlerini alabilirken, günümüzde ise bir konu hakkında milyonlarca görüşün (her görüşe ayrı ayrı bakmaya gerek kalmaksızın) toplam sonucuna (örneğin a konusuna bakış açısı %65 olumludur) bakarak bir konu hakkında çok daha gerçekçi bir sonuca varabilmekteyiz.

Duygu analizi neden önemlidir?

Dünyadaki verilerin %80 inin yapılandırılmamış olduğu, var olan verilerin %90'ının son iki yılda üretildiği (Schneider, 2016) ve önceden tanımlanmış bir şekilde düzenlenmediği tahmin edilmektedir. Bu verileri çoğu elektronik posta, forumlar, sohbetler, sosyal medya, anketler, makaleler ve belgeler gibi metin verilerinden gelmektedir. Bu metinleri analiz etmek, anlamak ve sınıflandırmak zor, zaman alıcı ve

maliyetli bir süreçtir. Duyarlılık analiz sistemleri bu işlemleri hızlı bir şekilde yapıp daha doğru sonuçlara ulaşmaya olanak sağlamaktadır.

Duygu analizi yapılırken kullanılabilecek üç temel yöntem vardır (Anonim, 2017c);

- Daha önce elle hazırlanmış kurallar kümesine dayalı analiz yapan "*Kural Tabanlı Sistemler*".
- Verilerin öğrenmesi için otomatik "*Makine Öğrenme*" tekniklerine dayanan sistemler.
- Hem kural tabanlı hem otomatik yaklaşımları birleştiren "*Hibrit Sistemler*".

Kural tabanlı sistemler, daha önce belirlenmiş kuralları, etiketleri, sözcük ve ifade listelerini içermektedir. Kelimeler sözlükte olumlu, olumsuz ve tarafsız (nötr) olarak tanımlanır. Analiz edilecek metinden tarafsız ifadeler çıkarılarak, negatif ve pozitif kelimelerin durumuna bakılarak metnin polaritesi hesaplanır. Bu sistemde kelimelerin nasıl birleştirildiği hesaba katılmadığı için düşük oranda da olsa hatalar barındırmaktadır. Kelime listesine sürekli yeni kelimeler ve ifadeler eklenmesine ihtiyaç duymaktadır. Bunun yanında analizde çok hızlı sonuçlar vermektedir.

Otomatik öğrenme sistemlerinde, kural tabanlı sistemlerin aksine bir sınıflandırıcının bir belirli bir metinle beslendiği (eğitim veri seti) makine öğrenme tekniklerine dayanmaktadır. Bu eğitim veri seti temel alınarak test veri setinde üzerinde analiz yapılmaktadır. Eğitim sürecinden sonra, yani testteki verilerin de polaritesi hesaplandıktan sonra, analiz edilecek metnin (verinin) sınıflandırmalar yapılarak hangi kategoriye ait olduğunu tahmin etmeye çalışmaktadır.

2. LİTERATÜR BİLDİRİŞLERİ

Bilgisayar ve yazılımların maliyetlerinin düşmesi ve kapasitelerinin artması sonucu kullanım alanı hayatımızın neredeyse her alanını kaplamaktadır. Verilerin baş döndürücü hızlarda artmasının sonucu veriyi analiz etmek ve bu veriden değerli bilgi elde etmek çok önemli bir yer edinmektedir. Verilerin devasa boyutlara ulaşmasının sonucunda hayatımıza yeni bir kavram olan "Büyük Veri-Big Data" girmiştir. Günümüzde büyük veri sadece bilişim alanında değil hayatımızın neredeyse her alanında kendini göstermektedir (Mayer-Schonberger, 2013).

Büyük verinin oluşumunda sadece bilişim ile i lgilenen büyük firmalar veya kurumlar değil bizler de sürekli veri üreterek katkı sunmaktayız (Dirin, 2017).

Büyük veri, web sayfalarının sunucu günlükleri (web server log), internet günlükleri (bloglar), cep telefonları, müşteri hizmetleri kayıtları gibi çeşitli kaynaklardan gelen çok miktardaki bilgiyi içermektedir (Snijders ve ark., 2012).

2000'li yıllarda hayatımıza girmeye başlayan büyük veri kavramı ilk tanımlanırken, o günün şartlarındaki veriler düşünülerek bir tanımlama yapılmıştır. Belli bir hızda büyüyüp belli bir büyüklüğe ulaşan ve bir çeşitliliğe sahip olan (3V) veriler büyük veri olarak tanımlanırken (Laney, 2001) günümüzde bu tanımın yetersiz geldiği, bilginin doğruluğunun (Verification) ve bilginin ne kadar değerli (Value) olduğuna dikkat çekilerek 5V olarak tanımlanmaktadır (Bisk, 2017; Biehn, 2017).

Günümüzde internetin yaygınlaşması ile birlikte kullanıcılar, 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 (Utkun, 2012).

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 aşamada veri özetleme, sınıflandırma, veriler arasındaki bağıntıların bulunması gibi farklı birçok teknik kullanılmaktadır (Baykal, 2006).

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

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 yapmakta veya yapmayı planlamaktadır (Snijders, 2017).

Artan veriler üzerinde sınıflandırma, özetleme, büyük veriden küçük ve anlamlı sonuçlar elde edebilmek için "veri madenciliği" alanında çalışma yapmak gerekmektedir (Savaş ve ark., 2012). Çünkü artık geleneksek yöntemler (ilişkisel veri tabanı yönetim sistemleri) yetersiz kalmaktadır. Elde var olan veriler arasındaki ilişkiyi tanımlamak ve o verilerden yola çıkarak sonraki verilerin nasıl olacağını tahmin etmek için "Veri Madenciliği" kavramı hayatımıza girmiştir (Aydın, 2007).

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 (Davenport, 2012).

Seattle ve Los Angeles eyaletlerinde "Önleyici Polis Hizmetleri" adında bir uygulama hayata geçirilmiş, işlenen suçların yeri, zamanı ve içeriği incelenmiştir. Dört aylık bir süreçte cinayet oranının %12 ve hırsızlığın %26 azaldığı gözlemlenmiştir (Hvistendahl, 2016).

Hollanda istatistik kurumu, yollarda bulunan algılayıcılardan elde edilen verilerden yola çıkarak, yolların kullanım oranlarını tespit etmiş, her aracın hızını ve tipini (minibüs, taksi, kamyon, kamyonet vb.) belirleyerek merkezi veri tabanına bildirmiştir. Bu çalışmayla ulaşımda alınması gereken önlemler belirlenmiştir. Projede toplanan veri boyutunun yüksekliğinden dolayı büyük veri araçları tercih edilmiştir (Domenico, 2013).

PriceStats, çevrimiçi fiyatlardan faydalanarak enflasyon tahmini yapan bir yazılım olup bu yazılım, doktora öğrencisi Alberto Cavallo tarafından doktora tezi kapsamında

geliştirilen bir projedir (Cavallo,2017). 2012 yılında "The Economist" dergisi, Arjantin'in, resmi enflasyon rakamları yerine bu projeden faydalanarak bir sonuç elde etmiştir. Ekonomistler, PriceStats'ın hesapladığı rakamların, ülkelerin resmi istatistiklerinden daha güvenilir bulmaktadır. 2014 yılında ülkemiz de proje kapsamına alınmıştır.

Web'in hayatımıza girmesiyle üretilen veri artık sadece metin veriler değil fotoğraf, video ve ses dosyaları da veri olarak üretilmeye başlanmıştır(Getting, 2017, Peter, 2010). Özellikle Web 2.0'dan (Okunabilir/Yazılabilir Web) gelişmesiyle beraber artık internete bağlı her kullanıcı veri üretmeye başlamıştır (Karabulut, 2009). Üretilen verilerin anlık ve çift yönlü paylaşımına olanak tanınması ile "Sosyal Medya" kavramı hayatımıza girmiştir (Çelik, 2018).

Günümüzde daha çok insanın internete ulaşması ve sosyal medyayı kullanımlarının artmasına paralel olarak üretilen veriler de büyümektedir ve önemi de aynı ölçüde artmaktadır. Büyük verinin 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 (James, 2016);

- 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.

2006 yılında Jack Dorsey tarafından kurulan ve başlarda sadece iş arkadaşları ile durumu paylaşma temeline dayanan (stat.us - durumumuz) Twitter, daha sonra diğer insanların kullanımına açılarak başlarda sadece gereksiz bilgiler ve kuş cıvıltıları olarak görülmüş (Sarno, 2009) ve 2017 yılının başlarında 328 milyon aktif kullanıcıya ulaşmıştır (Kwak, 2010).

Günümüzde Twitter ciddi bir kullanıcı sayısında sahip, veri tabanı açık ve verileri üzerinde çalışma yapılabilmesine olanak sağlayan iyi bir veri kaynağı haline gelmiştir (Poynter, 2012).

Veri madenciliği için önemli bir kaynak olmasına dayanarak, Twitter üzerinden "kripto para", "yapay zeka" ve "makine öğrenimi" olmak üzere üç ana kategoride veriler elde edilmiş ve analizleri yapılmıştır.

Makine öğrenimi, verilerden yola çıkarak öğrenimi olanaklı kılan veya hedefleyen bir alandır. Makine öğreniminin amacı, bilgisayarlara veriler arasındaki ilişkiyi tanımlama veya gelecek verilerin nasıl olacağı ile ilgili tahmin yürütmektir (Anonim, 2017b).

Kripto para, kriptografi bilimini kullanarak geliştirilen bir para birimi olup bankacılık sektöründen ve para otoritelerinden sıyrılarak çok düşük işlem maliyetlerine sahip finansal işlemlere olanak sağlamaktadır (Hayes, 2017).

Duyarlılık analizi veya fikir madenciliği, bir metin üzerinde çalışarak, o metnin içinde "Doğal Dil İşleme" (NLP- Natural Language Processing) algoritmaları ile bir görüş belirlemeye çalışmaktadır. Bu, insanların belli bir konuya eğilimlerini tespit etmek için önemlidir (Pang, 2008).

3. METARYAL VE YÖNTEM

3.1. Materyal

Bu çalışmada Linux işletim sisteminin Ubuntu dağıtımı kullanılarak, Python programlama dili ile Twitter verileri elde edilmiştir. Twitter uygulamasından veri çekebilmek için Twitter API'den faydalanılmış, Python tarafında ise API ile sorunsuz iletişim sağlamak için "tweepy" kütüphanesi kullanılmıştır. Veriler elde edildikten sonra üzerinde yapılan işleme göre CSV formatında ve MySQL veri tabanına kaydedilmiştir. Bu veriler üzerinde temizleme, sınıflandırma ve sayısal işlemler için Python programlama diline ait "pandas", "textblob", "scikitlearn" ve görsel sonuçlar için "matplotlib" kütüphaneleri kullanılmıştır. Aşağıda bu kavramlar ve uygulamanın işleyişi hakkında ayrıntılı bilgi verilmiştir.

3.1.1. Ubuntu

Linux, Unix çekirdeği baz alınarak geliştirilen, açık kaynak kodlu ve destekler hariç ücretsiz bir işletim sistemi çekirdeğidir. Çekirdeğin kaynak kodları "Genel Kamu Lisansı" (GNU) çerçevesinde değiştirilebilir, geliştirilebilir veya dağıtılabilir.

Linux ismi, 1991 yılında yazarı olan Linus Torvalds ve Unix'e de atıfta bulunarak geliştirilen Linux çekirdeğinden gelmektedir. Ancak buna rağmen Linux'un Unix ile herhangi bir kod ortaklığı bulunmamaktadır. Diğer bir deyişle Linux'un kodları baştan sonra yazılmıştır. Linux çekirdeği; geniş bir donanım desteğine sahip olmasından dolayı da Sunucu bilgisayarlarda, masaüstü-dizüstü bilgisayarlarda, iş istasyonlarında, tabletler ve akıllı telefonlar gibi hemen her platformda uyum içerisinde çalışabilmektedir. Kullanım oranı bakımından Linux, sunucu işletim sistemlerinde ilk sırada tercih edilmektedir. Aynı zamanda ve dünyanın 10 hızlı süper bilgisayarında kullanılmaktadır.

Linux; işletim sistemi değildir, işletim sistemi çekirdeğidir. Linux çekirdeği, bu çekirdeği kullanan birçok Linux dağıtımı tarafından kullanılarak bir işletim sisteminin sağladığı bütün özellikleri sağlamaktadır. Bir Linux dağıtımı; Linux çekirdeği ve grafik ara yüzünün bir araya gelmesiyle, yönetimsel yapılandırma araçları seti, yazılım

güncelleme araçları vb. ile oluşturulan bir tam özellikli bir işletim sistemini ifade etmektedir.

Günümüzde Linux çekirdeğini kullanan dağıtımları "Linux" adıyla anılmaktadır. Linux işletim sisteminin her aşaması açık olarak internet üzerinde yayınlanmakta, bu işletim sistemlerini kullanan kişiler tarafından test edilmekte, hataları ve eksiklikleri belirlenerek düzeltilmekte ve geliştirilmektedir. Hatalar, anında kullanıcılar tarafından belirlenip rapor edilmekte ve birçok kişinin katkısıyla düzeltilmektedir (Anonim, 2017g).

İlk kararlı sürümü 2004 yılında yayınlanan Ubuntu, "Canonical Limited" firması tarafından, Linux çekirdeği temel alınarak geliştirilen, açık kaynak kodlu bir işletim sistemidir. Masaüstü, sunucu, tablet, nesnelerin interneti gibi birçok ihtiyaca yönelik geliştirilen türevleri bulunmaktadır. Altı ayda bir yeni sürümü yayınlanmaktadır, iki yılda bir ise LTS (Long Term Support- Uzun süreli Destek) sürümleri yayınlanmaktadır. LTS sürümlerine beş yıl boyunca sürüm yükseltme gerektirmeden destek verilirken, ara sürümlere 9 ay destek verilmektedir.

Ubuntu'nun geliştiricisi Canonical Limited, masaüstü, sunucu, bulut ve nesnelerin internetine (IoT- Internet of Things) yönelik olarak "Ubuntu Desktop", "Ubuntu Server", "Ubuntu Cloud", "Ubuntu Core" isimli türevlerini resmi olarak geliştirmekte ve desteklemektedir (Anonim, 2017e).

3.1.2. Twitter API

İnternetin yaygınlaşması ile internet ortamında çalışan uygulamalar kendi içinde büyük boyutlarda veri depolaması yapabilmektedir. Bunun yanında internet uygulamaları kendi aralarında ve üçüncü şahıslara veri paylaşımı ihtiyacı duymaktadır. Bir uygulama diğer bir uygulamaya veya üçüncü şahıslara veri paylaşımını genellikle kendi geliştirdiği API'ler (Application Programing Interface – Uygulama Programlama Arayüzü) yardımı ile sağlamaktadır.

Twitter, sunucularında toplanan bütün verilere kendisi tarafından geliştirilen "Twitter API" sayesinde dışarıdan erişimlere olanak sağlamaktadır. Dışarıdan gelen taleplere cevap olabilmek adına sürekli geliştirilen Twitter API, yazılım geliştiriciler,

diğer internet uygulamaları ve veri analistleri için iyi bir kaynak olmaya devam etmektedir (Anonim, 2017c).

Twitter Streaming API, HTTP protokolünü kullanarak; POST, GET ve DELETE isteklerine gerçek zamanlı olarak cevap verebilmektedir (Bifet, 2010). Başka bir deyişle, hesap sahibi gerçek zamanlı olarak kendi tweetlerini okuyabilmekte, değiştirebilmekte veya silebilmektedir. Kendi hesabı dışındaki diğer açık (public) hesapların tweetlerini ise sadece okuyabilmektedir.

3.1.3. Python

Python, "Guido Van Rossum" isminde Hollandalı bir programcı tarafından geliştirilen programlama dilidir. Geliştirilmesine 1990 yılında başlanan Python; karşılaştırırıldığı diğer dillere (C ve C++) kıyasla daha kolay öğrenilir, daha hızlı ve sade yazılabilir, bir derleyiciye ihtiyaç duymadan daha sade bir söz dizimine sahip bir dil olarak karşımıza çıkmaktadır. Bu özelliklerinden dolayı Google, Yahoo ve Dropbox gibi dünyanın en büyük yazılım firmaları bünyelerinde her zaman Python programcıları bulundurmaya ihtiyaç duymaktadır. Bilinenin aksine Python bir yılan isminden değil, geliştiricisinin "Monty Python's Flying Circus" isimli televizyon dizisini çok sevmesinden gelmektedir (Yaman, 2016).

Son yıllarda bilimsel işlemlerdeki hızından dolayı bilim dünyasında ağırlıklı olarak kullanılmakta, özellikle metin sınıflandırma, grafik makine öğrenmesi gibi işlemlerde kullanımı hızla artmaktadır. Hızı kadar geniş bir kütüphane desteği sayesinde her geçen gün daha fazla tercih edilmektedir. Google arama motoruna yazılan bir kelimenin saniyeler içinde milyonlarca sonuç getirmesi Python sayesinde mümkün olmaktadır. Bilim ve teknoloji alanında kullanımının yansıra askeri savunma alanında, hacking ve network işlemlerinde de kullanımı artmaktadır.

Yukarıda anlatıldığı gibi Python, özellikle geniş kütüphane desteği sayesinde birçok işlemi kolaylıkla yapabilmektedir. Web geliştiricilerinin kullandığı Django, Pyramid, Flask ve Bootle gibi frameworkler Python ile yazılmıştır. Ayrıca standart kütüphaneleri birçok internet protokolünü desteklemektedir. HTML(HyperText Markup Language), XML(Extensible Markup Language) protokollerini desteklemenin yanında, e-mail işlemleri, FTP (file Tranfser Protocol), IMAP (Internet Message Access Protocol) ve diğer internet protokollerini desteklemektedir (Anonim, 2017f).

Tezin uygulamasında kullanılan Python kütüphaneleri aşağıda gösterilmiştir.

Tweepy. Twitter Streaming API ile belli parametrelerle bağlantı kurarak verilerin JSON formatında elde edilmesi için kullanılmaktadır.

TextBlob. Metin verilerini işlemek için bir Python kütüphanesidir. Konuşma etiketleme, isim cümlesi çıkarma, duygu analizi, sınıflandırma, çeviri ve daha fazlası gibi ortak doğal dil işleme (NLP) görevleri yerine getirmektedir.

Pandas. Python programlama dili için geliştirilen yüksek performanslı, kullanımı kolay, veri yapıları üzerinde analiz araçları sağlayan açık kaynak kodlu bir kütüphanedir.

Mysql.Connector. Tweepy ile elde edilen verileri MySQL veritabanına ekleyen ve veriler üzerinde işlemler yapmaya olanak sağlayan kütüphanedir.

Re (Regular Expression. Düzenli ifade). Elde edilen tweetleri gereksiz karakter ve içeriklerden arındırılmasına olanak sağlamaktadır. Başka bir ifade ile metinleri temizlemektedir.

MatPlotlib. Sayısal verilerin grafik ortamına aktarılmasında kullanılmaktadır. Yani verileri görselleştirmektedir.

NumPy. Bilimsel hesaplamalarda kullanılan Python kütüphanesidir.

ScikitLearn (sklearn). Makine öğrenmesinde kullanılan bu kütüphane, genel olarak belli bir veriyi örnek alarak sonraki verilerin özelliklerini tahmin etmeye çalışır. Metinler üzerinde sınıflandırma, eğitim ve veri setleri oluşturma, gelen verinin hangi sınıfa ait olduğunu tahmin etme ve bu işlemleri görselleştirmek üzere sayısallaştırma gibi işlemleri gerçekleştirmektedir.

3.1.4. Metin sınıflandırma algoritmaları

Otomatik metin sınıflandırma veya metin kategorizasyonu makine öğrenmede önemli bir alt başlık olup elektronik ortamdaki metin verilerin çok hızlı artması ile de her geçen gün daha da önem kazanmaktadır. Üzerinde çalışılan veri tiplerinin metin olmasından dolayı bilimsel çalışmalarda sık kullanılan metin sınıflandırma algoritmaları kullanılmış, bu algoritmaların doğruluk skorları ve bu skorları üretme süreleri hesaplanıp karşılaştırılmıştır.

3.1.4.1. Naif bayes sınıflandırıcı

Naif Bayes sınıflandırma algoritması, 1763 yılında geliştirilen, adını Matematikçi Thomas Bayes'den alan, metin veriler üzerinde çalışan bir sınıflandırma ve kategorilendirme algoritmasıdır. Naïve Bayes, "Bayes Teoremi"ni dayanan olasılıksal bir sınıflandırıcıdır. Daha basit bir ifade ile metin girdilerini sınıflandırırken, sınıfın nitelikleri arasında belli bir bağlantı olmadığı varsayımına dayanmaktadır (Ceci, 2005).

Son yıllarda verilerin artması nedeniyle Bayes teoreminin kullanımı da aynı şekilde artmaktadır. Temeli her ne kadar Thomas Bayes tarafından ortaya atılmış olsa da sonraki yıllarda Finette Savage ve Linette gibi isimler bu teoremin gelişimine katkıda bulunmuşlardır (Jatana & Sharma, 2014).

Naif Bayes yaklaşımı, maksimum olasılık ilkesine dayanan, olasılıkları hesaplamak için kullanılan iki rastgele olayın koşullu olasılıklarını hesaplamaya çalışan bir teoremdir. Bu durumda, Bayes Teoremi, rastgele bir olay elde etmek için A ve B olasılıklarının doğruluğunu hesaplamak için kullanılabilir. Naif Bayes sınıflandırması, en yüksek olasılığa sahip sınıfın örneklerinin olasılık sonucunu tahmin etmeye çalışır. A ve B rastgele ve bağımsız olduğu varsayımına dayanır. Bayes teoreminin denklemi Eş. 3.1'e gösterilmiştir.

$$P(A|B) = \frac{P(B|A)\ P(B)}{P(A)} \tag{3.1}$$

P(A): A olayının bağımsız olasılığı

P(B): B olayının bağımsız olasılığı

P(B|A): *A* olayının bilinmesi (gerçekleşmesi) durumunda *B* olayının meydana gelme olasılığı

P(A|B): *B* olayının bilinmesi (gerçekleşmesi) durumunda *A* olayının meydana gelme olasılığı

Bir örnek ile açıklamak gerekirse, örneğin dünyada işlemci üreten A ve B firmaları olsun. A firmasının işlemci üretim oranı %30 ve B firmasının işlemci üretim oranı %70 olsun. Ayrıca A firması ürettiği işlemcilerin defolu olma olasılığı %4 ve B firmasının %6 olsun. Bu olasılıkların gösterim şekli sırasıyla aşağıda gösterilmiştir.

P(A): 0.3

P(B): 0.7

P(Defolu|A): 0.04
P(Defolu|B): 0.06

Rasgele seçilen bir işlemcinin defolu olduğu bilindiğinde A firmasından olma olasılığı Bayes teoremine göre hesaplanması Eş. 3.2'de gösterilmiştir.

$$P(A|Defolu) = \frac{P(defolu|A) * P(A)}{P(defolu)}$$
$$= \frac{0.04 * 0.3}{0.04 * 0.3 + 0.06 * 0.7}$$
$$= \frac{0.012}{0.012+0.042} = 0.22 \quad (3.2)$$

Burada bir ürünün defolu olma olasılığı Eş. 3.3'te gösterilmiştir.

$$P(defolu) = P(A) * P(Defolu|A) + P(B) * P(Defolu|B) \quad (3.3)$$

Naif Bayes sınıflandırmanın daha anlaşılır olması için bir örnek ile açıklamak gerekirse; elimizde tweetlerden oluşan bir veri seti olsun. Ve bu tweetleri sınıflandırmaya çalışalım. Bir tweetin hangi sınıfa ait olduğunu hesaplamak için kullanılan denklem Eş. 3.4'te gösterilmiştir.

$$P(C_p \mid T) = \frac{P(T|C_p)\, P(C_p)}{P(T)} \quad (3.4)$$

C_p: Olumlu olan tweet sınıfını temsil eder.
T: $\{ t_1, t_2, t_3, t_4 \ldots t_n \}$ şeklindeki tweetleri temsil eder.

3.1.4.2. Multinomial naif bayes sınıflandırıcı

Multinomial sınıflandırma tekniği temelde Bayes teoremini baz aldığından Python programlama dilinde Naif Bayes kütüphanesinin altında yer almaktadır ve programdaki kullanımı MultinomialNB olarak geçmektedir. Temel işlevi, yeni ele alınan metnin, en yüksek olasılıkla hangi sınıfa ait olduğunu tahmin etmeye çalışmaktır (Kibriya

ve ark., 2014). Bu tahmini yürütürken Eş. 3.5'e göre metinleri(tweetleri) sınıflandırmaya çalışır.

$$P(t_i|c) = (\sum_n f_{ni})! \prod_n \frac{P(w_n)^{f_{ni}}}{f_{ni}!} \quad (3.5)$$

t_i: Metinleri (tweetleri) temsil eder.

f_{ni}: Metinlerimizdeki kelime sayısını temsil eder.

c: kelimelerin veya metinlerin ait olduğu sınıfı (class) temsil eder.

$P(w_n)$: n kelimesinin c sınıfına ait olma olasılığını temsil eder,

3.1.4.3. Bernoulli modeli

Bernoulli modeli, bir belgedeki kelime uzayını ikili vektör olarak ele almaktadır. Burada üzerinde işlem yapılacak kelimeleri *V*, kelime uzayındaki her bir boyutu t ($t \in \{1, \dots, |V|\}$) ve sözlükteki her bir kelime için w_t gösterimleri ifade etmek üzere d_i belgesinin bir boyutunu temsil eden *t,* ve her bir kelimenin sözlükte olup olmadığının ifadesi de B_{it} olsun. Böyle bir belge temsili ile Naif Bayes teoremini uygularsak. bir belgede var olan bir kelimenin olma olasılığı başka bir belgede olma olasılığından bağımsız olarak görülmelidir. Bu varsayımla belgedeki kelimelerin sınıflandırılması için Eş. 3.6'da gösterilen deklem kullanılmaktadır (McCallum, 1998).

$$P(d_i|c_j\,;\,\theta) = \prod_{t=1}^{|V|}(B_{it}\,P(w_i|c_j\,;\,\theta) + (1 - B_{it})(1 - P(w_i|c_j\,;\,\theta))) \quad (3.6)$$

Bu nedenle, bir oluşturma bileşeni verildiğinde, bir belge çoklu bağımsız Bernoulli deneylerinin bir koleksiyonu olarak görülebilir. Her bir kelime ve her bir belgenin bağımsız olasılık teoremi ile hesaplanmasının ifadesi de denklemdeki $P(w_i|c_j\,;\,\theta)$ ile gösterilmiştir.

3.1.4.4. Ridge regresyon

Ridge regresyon, çoklu doğrusal olan metinler üzerinde kullanılan bir analiz tekniğidir. Yani değişkenler (metin veya kelimeler) arasında birden fazla bağlantı olduğunda, sınıflandırma tahmini yapmak daha zordur. Ridge regresyon tekniği, veri setindeki metinleri sınıflandırmaya çalışırken, biraz yanlılık ekleyerek hata oranını düşürmeyi amaçlamaktadır. Bu nedenle yanlı tahmin edici olarak da isimlendirilmektedir.

Ridge regresyon tahmin edicisi X^TX matrisi şeklindeki verilerin her bir köşegen elemanına sabit bir katsayı (k) ekleyerek hata paylarını azaltmak amacı ile Hoerl ve Kennard tarafından 1970 yılında önerilmiştir (Vinod ve Ulah, 1981). Yanlılık eklenerek hesaplanan regresyon analizi Eş. 3.7'de gösterilmiştir.

$$\beta(k) = (X^TX)^{-1}X^Ty, \qquad k \geq 0, \tag{3.7}$$

şeklinde ifade edilir. Burada $k \geq 0$ yanlılık parametresini ifade etmektedir.

3.1.4.5. Lojistik regresyon

Lojistik regresyon analizinde temel amaç, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi en az değişken ile en iyi uyuma sahip olacak bir model kurmaktır (Atasoy, 2001). Diğer regresyon analizlerinde olduğu gibi bir tahmin edicidir. Verileri tanımlamak ve bağımlı iki değişken ile bir veya daha fazla bağımsız değişkenler arasındaki ilişkiyi açıklamaya çalışmaktadır. Sonuç sadece ikili (1 veya 0) bir değişken ile ölçülmektedir.

Lojistik regresyon metin sınıflandırma ve doğal dil işleme alanlarında en çok kullanılan yöntemlerden biridir. Varsayılan sınıfın olasılığını ölçmektedir. Mesela bir metnin (tweet'in) olumlu olma olasılığını hesaplamak için kullanılabilir. Koşullu olasılıkları Eş. 3.8'deki gibi tanımlar (Yu ve ark., 2011).

$$P_w(y = \pm 1|x) = \frac{1}{1+e^{-yw^Tx}} \tag{3.8}$$

x: işlenecek veri.

$y \in \{1, -1\}$: tahmin edilecek sınıf (olumlu veya olumsuz).

$w \in R^n$: Ağırlık vektörü.

Lojistik regresyon, Eş. 3.9'da gösterilen şekilde normalleştirilmiş negatif (veya pozitif) olabilirliğini hesaplamaktadır. *C*, bir sınıfa ait olma durumunu göstermektedir.

$$P^{LR}(w) = C \sum_{i=1}^{l} \log\left(1 + e^{-y_i w^T x_i}\right) + \frac{1}{2} w^T w \quad (3.9)$$

3.1.4.6. Doğrusal destek vektör sınıflandırma (linear support vector classification)

Doğrusal SVM, lineer destek vektör makinesi tasarımı için bir kesme düzlemi algoritmasının orijinal tescilli versiyonunu uygulayan çok büyük veri setlerinden çok sınıflı sınıflandırma problemlerini çözmek için kullanılan son derece hızlı makine öğrenimi (veri madenciliği) algoritmasıdır (Anonim, 2014). Metin sınıflandırma gibi birçok uygulamada büyük ölçekli sınıflandırma problemlerini çözmek çok önemlidir. Doğrusal sınıflandırma, çok sayıda örneği ve özelliği olan büyük seyrek veriler için en umut verici öğrenme tekniklerinden biri haline gelmiştir.

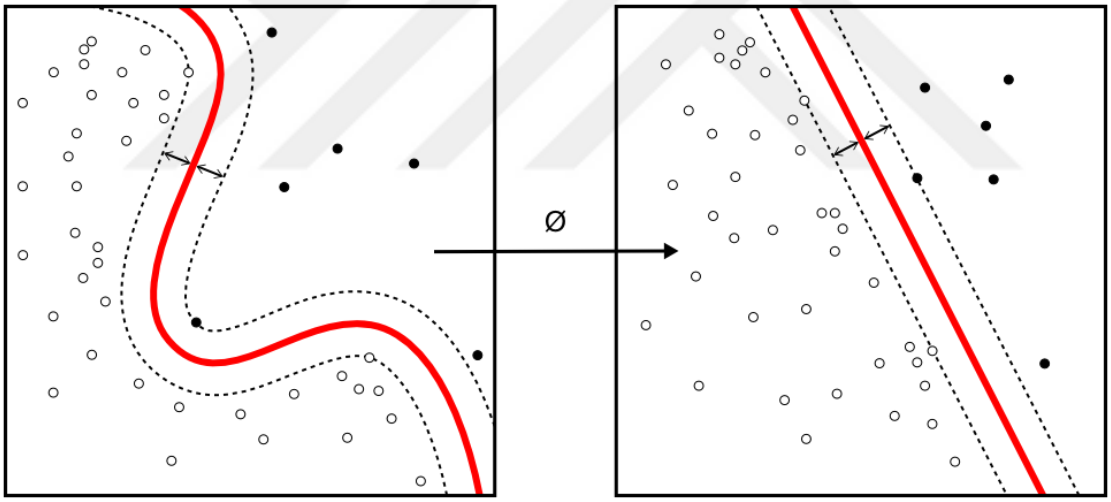
Destek vektör makineleri (SVM'ler) veri sınıflandırması için popüler olmuştur. Bir SVM genellikle verileri yüksek boyutlu bir alana eşler ve daha sonra çekirdek tekniklerini kullanır. Doğrusal olmayan SVM eğitimi genellikle popüler metin ayrıştırma metotları kullanılarak gerçekleştirilir. Bununla birlikte, bu ayrıştırma yaklaşımları, büyük veri kümeleri için çok zaman gerektirir. Ek olarak, test prosedürleri, destek vektörlerini ve test örneklerini içeren çekirdek hesaplaması nedeniyle yavaştır (Cortes ve Vapnik, 1995).

Daha genel bir ifadeyle destek vektör makinesi, sınıflandırma, gerileme veya aykırı saptama gibi diğer görevler için kullanılabilen, yüksek veya sonsuz boyutlu verilerde çok değişkenli veya çok sınıflı model kümesi oluşturur. Sezgisel olarak, herhangi bir sınıfın en yakın eğitim veri noktasına en yakın mesafeye sahip olan hiper düzlem tarafından iyi bir ayrılma elde edilir (sözde işlev marjı olarak adlandırılır), çünkü genel olarak marj ne kadar büyükse sınıflandırıcının genelleme hatası o kadar düşüktür (Anonim, 2016). Makine öğrenimi ve istatistiksel öğrenme teorisindeki denetimli

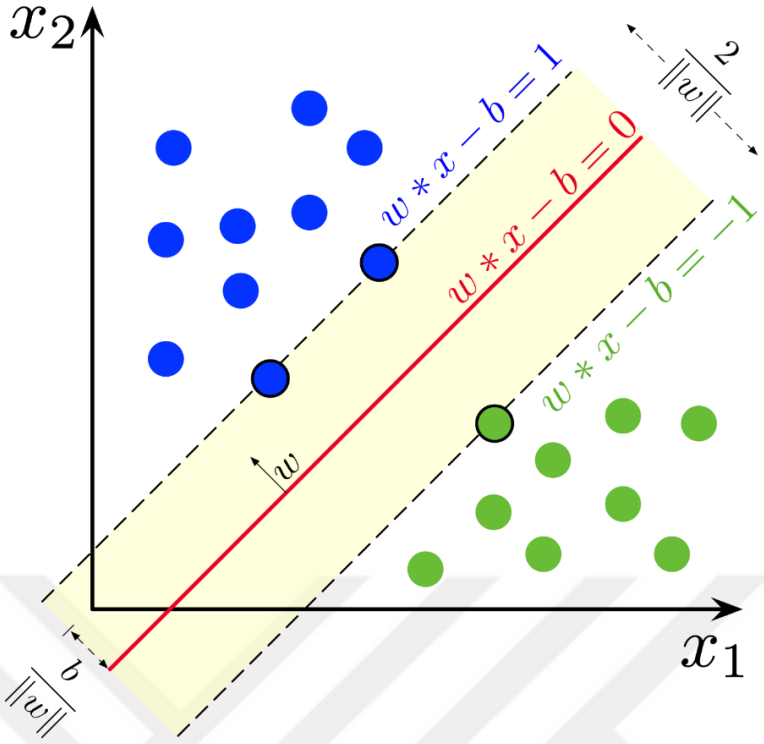
öğrenme uygulamalarında, genelleme hatası, bir algoritmanın daha önce görülmemiş veriler için sonuç değerlerini ne kadar doğru tahmin edebildiğinin bir ölçüsüdür. Öğrenme algoritmaları sonlu örneklerde değerlendirildiğinden, bir öğrenme algoritmasının değerlendirilmesi örnekleme hatasına duyarlı olabilir.

Destek vektör makineleri, sınıflandırma, yoğunluk tahmini ve kümeleme için kullanılır. Doğrusal olmayan veri setini, şekil 3.1'de gösterildiği gibi doğrusal olarak ayrılabileceği yüksek boyuta aktarıp veriler arasındaki en yüksek sınıfın bulunması temeline dayanır (Elmas, 2012).

Verileri doğrusal ayırabilen SVM, Şekil 3.1'de gösterildiği gibi verileri ayıran bir hiper düzlem çizer. Ancak SVM'lerin asıl hedefi, kendisine en yakın noktalar arasındaki uzaklığı en yüksek seviyeye çıkararak en uygun hiper düzlemi bulmaktır. Sınırı en yükseğe çıkardıktan sonra en uygun ayırımı yapan düzleme optimum düzlem (Şekil 3.2), ve sınırları gösteren noktalara ise destek vektörleri denilmektedir (Yelmen, 2016).



Şekil 3.1. Doğrusala çevirme (Anonim, 2016).



Şekil 3.2. Destek vektörleri ve optimum hiper düzlem (Anonim, 2016).

3.1.4.7. Pasif agresif sınıflandırma (Passive agresive classifier)

Pasif Agresif Sınıflandırıcı, Crammer tarafından yayınlanan bir makine öğrenimi algoritmasıdır. Hem sınıflandırma hem regresyon için kullanılabilir. Bu algoritmaya göre yeni gelen veriler (tweetler), eski veriler ile aynıysa, yeniden sınıflandırma için bir hesaplama yapmaz, eski veriye göre otomatik olarak sınıflandırır, fakat yeni veri farklıysa, sınıflandırma için yeniden hesaplama yapar (Anonim, 2018c). Tezin uygulamasında kullanıldığı şekliyle, "X" tweetleri "Y" polariteyi temsil ediyor olsun (Eş. 3.10). "X" değerlerini sınıflandırmak için kullanılan matematiksel eşitlik Eş. 3.11'deki gibi hesaplama yapmaktadır.

$$\begin{cases} X = \{x_0, x_1, x_2, x_3 \dots x_t \dots .\}\ x_i \in \mathbb{R}^n \\ Y = \{y_0, y_1, y_2, y_3 \dots y_t \dots .\}\ y_i \in \{-1, +1\} \end{cases} \tag{3.10}$$

$$w_{t+1} = w_t + \frac{\max(0, 1 - y_t(w^T . x_t))}{\|x_t\|^2 + \frac{1}{2C}} y_t x_t \tag{3.11}$$

3.1.5. TF-IDF

TF-IDF (Term Frequency – Inverse Document Frequency, Terim Frekansı – Ters Metin Frekansı), metin madenciliği olarak da anılan doğal dil işleme (NLP) konularında metinler üzerinde istatistiksel incelemeler yapan basit ve etkili bir yöntemdir (Ramos, 2003). TF-IDF, Eş. 3.12'da görüldüğü üzere, temelde metnin içinde bir kelimenin frekansını (kaç defa tekrarlandığını, TF) ve bir kelimenin bütün dokümanlarda kaç defa geçtiğinin ters logaritmasını (IDF) alarak hesaplar. Sonra ikisinin çarpımı, bir kelimenin bütün dokümanlardaki TF-IDF değerini verir.

$$w_{i,d} = tf_{i,d} * \log(\frac{n}{df_i}) \tag{3.12}$$

i: Kelimenin kendisi (terim)

d: Dokümandaki terim sayısı

$tf_{i,d}$: i teriminin d dokümanındaki frekansını temsil eder.

Bir kelimenin TF'si (terim frekansı), bir belgede o kelimenin ne kadar geçtiğini ifade eder. Örneğin yüz kelimelik bir belgede 15 defa kedi kelimesi geçmişse Eş. 3.13'deki gibi hesaplama yapılmaktadır.

$$TF_{kedi} = \frac{15}{100} = 0.15 \tag{3.13}$$

Bir kelimenin IDF'si ise (ters belge sıklığı), bu kelimenin bütün belgeler içinde ne kadar önemli olduğunun ölçüsüdür. Örneğin webde veya bir bilgisayarda 10 milyon belge olsun. Kedi kelimesinin geçtiği belge sayısı ise 300 bin olsun. Kedi kelimesinin bu 10 milyon belge için IDF hesaplaması Eş. 3.14'deki gibi yapılmaktadır. Kedi kelimesinin TF-IDF değeri ise Eş. 3.15'teki gibi hesaplanmaktadır.

$$IDF_{kedi} = \log\left(\frac{10000000}{300.000}\right) = \log(33.33) = 1.52 \tag{3.14}$$

$$W_{kedi} = TF_{kedi} * IDF_{kedi} = 1.52 * 0.15 = 0.228 \tag{3.15}$$

Bir kelimenin TF, IDF ve TF-IDF değerlerini Python koduyla hesaplanması aşağıda gösterilmiştir (Şekil 3.3, Şekil 3.4, Şekil 3.5).

```
def TF_Hesapla(wordDict,bow):
    tfDict={}
    bowcount = len(bow)

    for word, count in wordDict.items():
        tfDict[word] = count/float(bowcount)
    return tfDict
```

Şekil 3.3. Bir kelimenin TF hesaplaması.

```
def IDF_hesapla(docList):
    import math
    idfDict = {}
    N = len(docList)

    idfDict = dict.fromkeys(docList[0].keys(),0)

    for doc in docList:
        for word,val in doc.items():
            if val > 0 :
                idfDict[word] += 1

    for word,val in idfDict.items():
        idfDict[word] = math.log10( N / float(val))

    return idfDict
```

Şekil 3.4. Bir kelimenin IDF hesaplaması.

```
def TF_IDF_Hesapla(tfBow, idfs):
    tfidf = {}

    for word, val in tfBow.items():
        tfidf[word] = val * idfs[word]

    return tfidf
```

Şekil 3.5. Bir kelimenin TF-IDF hesaplaması.

Ancak bunun yanında TF-IDF hesaplaması diğer birçok parametresi ile beraber “scikit-learn” kütüphanesinde mevcuttur. Şekil 3,6’daki gibi yazılan Python kodunun çıktısı Şekil 3.7 deki gibi olur.

```
from sklearn.feature_extraction.text import TfidfVectorizer
belge = ["This is very strange",
         "This is very nice"]
vectorizer = TfidfVectorizer(min_df=1)
X = vectorizer.fit_transform(belge)
idf = vectorizer.idf_
print dict(zip(vectorizer.get_feature_names(), idf))
```

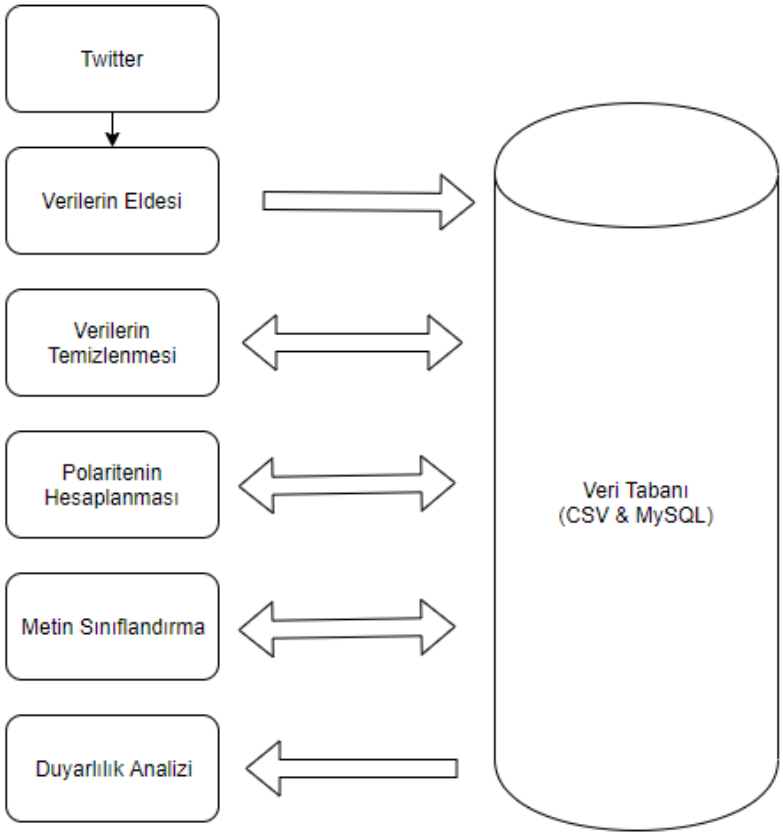
Şekil 3.6. Scikit-Learn ile TF-IDF hesaplaması. Python kodu.

```
{u'is': 1.0,
 u'nice': 1.4054651081081644,
 u'strange': 1.4054651081081644,
 u'this': 1.0,
 u'very': 1.0}
```

Şekil 3.7. Scikit-Learn ile TF-IDF hesaplaması. Ekran çıktısı.

3.2. Yöntem

Bu bölümde, verilerin nasıl çekildiği, sonrasında bu veriler üzerinde yapılan işlemler ele alınmıştır. Aşağıdaki şekilde yapılan işlemlerin aşamaları gösterilmiştir (Şekil 3.8).



Şekil 3.8. Uygulamanın işleyişi.

3.2.1. Verilerin eldesi

Tweetler elde edilirken, son yıllarda üzerinde en çok durulan konulardan kripto para, makine öğrenimi ve yapay zeka kategorilerinde tweetleri çekilmesi ve analizi hedeflenmiştir. Kripto paralardan en çok kullanılan ve işlem hacmi en yüksek olan para birimlerinden beş tanesine (Bitcoin, Ethereum, Bitcoin Cash, Ripple ve LiteCoin) ait 50.000'er adet, makine öğrenimine ve yapay zekaya ait 50.000 olmak üzere toplamda

300.000 adet tweet elde edilmiştir. 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 Twitter verilerini dinleyerek (son tarihten geriye doğru verileri çekerek) kaydedecek şekilde programlanmıştır.

Her bir konu çekilirken, o konuya ait anahtar kelimeler belirlenip (Çizelge 3.1) veri çekme işlemi ona göre gerçekleştirilmiştir.

Çizelge 3.1. Anahtar kelimeler

Para Birimi	Anahtar Kelimeler
Bitcoin	#BTC, #Bitcoin, Bitcoin
Ethereum	#ETH, #Ethereum, Ethereum
Bitcoin Cash	#BCH, #BitcoinCash, BitcoinCash
Litecoin	#LTC, #LiteCoin, LiteCoin
Ripple	#XRP, #Ripple, Ripple
Makine Öğrenimi	#MachineLearning, #DeepLearning, #ML,#DL
Yapay Zeka	#ArtificialIntelligence, #MachineIntelligence

Twitter Streaming API ile tweet elde edebilmek için öncelikle bir Twitter hesabına sahip olmak gerekmektedir. https://twitter.com adresinde kayıt işlemi yapıp bir Twitter hesabına sahip olduktan sonra, Twitter uygulamalarının yer aldığı https://apps.twitter.com/ adresinden veri elde edebilecek anahtar değerlerin oluşturulması gerekmektedir.

```
import tweepy
Consumer_key    = 'XXX'
Consumer_secret = 'XXX'
Access_token  = 'XXX'
Access_secret = 'XXX'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)

for tweet in tweepy.Cursor(api.user_timeline, screen_name="@m_c_erdogan"):
    print("ID   : " + str(tweet.id))
    print("Text : " + tweet.text)
    print("Tarih: " + tweet.created_ad)
```

Şekil 3.9. Tweepy ile oturum açma.

Şekil 3.9'da gösterilen Python kodunda "@m_c_erdogan" adlı kullanıcıya ait en sonra tarihten geriye doğru 200 adet tweete ait bilgileri ekrana yazar. Herhangi bir sayı belirtilmezse bu sayı varsayılan 200 tweettir. Verilerin sürekli çekilebilmesi için bu kod kısmının sürekli bir döngüde çalışması gerekmektedir (Şekil 3.10).

```
class listener(StreamListener):

    def on_data(self, data):
        all_data = json.loads(data)
        # tweetten istenilen alanlar elde edilebilir
        if 'text' in all_data:
          tweet             = all_data["text"]
          created_at        = all_data["created_at"]
          retweeted         = all_data["retweeted"]
          username          = all_data["user"]["screen_name"]
          user_location     = all_data["user"]["location"]

    def on_error(self, status):
        print(status + "Tweet Eklenemedi")
```

Şekil 3.10. Tweetlerin sürekli eldesi.

Burada elde edilen tweetlerin ekrana yazmak yerine MySQL veri tabanına ve CSV(comma-separated values- virgülle ayrılmış değerler) formatında kaydedilmektedir. Elde edilen tweetlerin MySQL veritabanına kaydetme kodları (Bkz. Şekil 3.11) ve CSV formatında kaydetme kodları gösterilmiştir (Bkz. Şekil 3.12).

```
import mysql.connector
from mysql.connector import errorcode
import time
import json

cnx = mysql.connector.connect(user='root', password='',
                              host='localhost',
                              database='dbname',
                              charset = 'utf8mb4')
cursor=cnx.cursor()

sql =  "INSERT INTO tweets (created_at, user, tweet,Location, retweeted)"
sql += "VALUES (%s,%s,%s,%s,%s,%s,%s)"
cursor.execute(sql,

    (created_at, user, tweet, location, retweeted))
cnx.commit()
```

Şekil 3.11. Mysql ‘e kaydetme.

```
import time
import json
import csv

WORDS = [ '#BTC', '#Bitcoin','bitcoin']
csvFile = open('allcrypto.csv', 'a')
csvWriter = csv.writer(csvFile)
csvWriter.writerow(['tweet_id','user','created_at','tweet','retweeted','location'])
class listener(StreamListener):

    def on_data(self, data):
        all_data = json.loads(data)

        if 'text' in all_data:

          tweet                 = all_data["text"]
          created_at            = all_data["created_at"]
          retweeted             = all_data["retweeted"]
          username              = all_data["user"]["screen_name"]
          user_location         = all_data["user"]["location"]
          tweet_id              = all_data['id']

          try:
             csvWriter.writerow([tweet_id,user,created_at,tweet,retweeted,location])
             print ('veri eklendi ', tweet_id)
          except:
            print('Hata Atlandi')

          return True
        else:
          return True

    def on_error(self, status):
        print(status)
```

Şekil 3.12. CSV formatında kaydetme.

Çizelge 3.2. Elde edilen örnek tweet tablosu

tweet_id	username	created_at	Tweet
105... 80	NokhezFatima	Tue Oct 30 12.23.18 +0000 2018	If You Are Hungry For More Than a Basic BTC Wallet, Then Bread Is For You! https.//t.co/9kdJfWFqzr #BTC #XRP #news
105... 76	CryptScy	Tue Oct 30 12.23.29 +0000 2018	Anybody notice the #jibehub, which @AmirSarhangi, @ripple's new hire, is CEO and co-founder of, looking a lot like□□ https.//t.co/rWqD4PMHw6
105...12	SmartPhonesTube	Tue Oct 30 12.23.38 +0000 2018	RT @CryptoSjop. Bitcoin kopen met iDEAL. Zo doe je dat bij de beste Bitcoin brokers https.//t.co/DZAQ9Bi3GP #HTMLCOIN #DGB #xvg $btc #IOTA□□
105...21	LarocheMatt	Tue Oct 30 12.23.40 +0000 2018	RT @sentosumosaba. Hey Hey Everybody ~ It's Eri in Tokyo, your update from Japan. Messaging App Leader + Messaging Payments Investment, XRP□□

3.2.2. Verilerin temizlenmesi

Tweetler üzerinde daha sağlıklı bir sınıflandırma ve duyarlılık analizi yapabilmek için, elde edilen tweetlerin temizlenmiş, yani gereksiz karakterlerden, linklerden, reklamlardan arındırılmıştır. Kısacası her tweetten "A-Z, a-z, 0-9" dışındaki karakterler çıkarıldıktan sonra veri tabanına kaydedilmiştir. Bunun için Şekil 3.13'te gösterilen kod parçası kullanılmış ve temizleme sonrası bir tweet örneği Çizelge 3.3'te gösterilmiştir.

```
import re

def clean_tweet(tweet):
    return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)",
                " ", tweet).split())
```

Şekil 3.13. Tweetlerin temizlenmesi.

Elde edilen tweetin temizlenmesi sonrası hali Çizelge 3.3'te gösterilmiştir.

Çizelge 3.3. Tweetlerin temizlenmesi

tweet	cleaned_tweet
If You Are Hungry For More Than a Basic BTC Wallet, Then Bread Is For You! https.//t.co/9kdJfWFqzr #BTC #XRP #news	If You Are Hungry For More Than a Basic BTC Wallet Then Bread Is For You BTC XRP news
Anybody notice the #jibehub, which @AmirSarhangi, @ripple's new hire, is CEO and co-founder of, looking a lot like□ https.//t.co/rWqD4PMHw6	Anybody notice the jibehub which s new hire is CEO and co founder of looking a lot like
RT @CryptoSjop. Bitcoin kopen met iDEAL. Zo doe je dat bij de beste Bitcoin brokers https.//t.co/DZAQ9Bi3GP	RT Bitcoin kopen met iDEAL Zo doe je dat bij de beste Bitcoin brokers
#HTMLCOIN #DGB #xvg $btc #IOTA□	HTMLCOIN DGB xvg btc IOTA
RT @sentosumosaba. Hey Hey Everybody ~ It's Eri in Tokyo, your update from Japan. Messaging App Leader + Messaging Payments Investment, XRP□	RT Hey Hey Everybody It s Eri in Tokyo your update from Japan Messaging App Leader Messaging Payments Investment XRP

3.2.3. Polaritenin hesaplanması

Polarite (kutupluluk), bir kelimenin tek başına "-1, 1" aralığında ne kadar olumlu veya olumuz olduğu anlamına gelmektedir. Veriler elde edilirken kaydedilmeden önce, her tweet elde edildiği anda, daha sonra ikinci bir işlem olarak yapılmaması ve sınıflandırmadan sonraki doğruluk skorlarında kullanılmak üzere polariteler hesaplanmıştır. Polariteler hesaplanırken, Doğal Dil İşleme (NLP) kapsamında yer alan "TextBlob" kütüphanesi içinde bulunan "sentiment.xml" adındaki, daha önce kelimelerin polaritesinin yazıldığı dosyadan yararlanılmıştır. Bu dosyanın içinde 2930 kelime kayıtlı olup yeni kelimeler eklenmeye devam edilmektedir. Şekil 3.14'te gösterildiği gibi bu dosyanın içinde sadece polarite oranı değil, bunun yanında kelimenin özniteliği (öznel mi nesnel mi), yoğunluğu (sonraki kelimeyi ne kadar etkilediği) ve güven değerleri de kaydedilmiştir. "sense" anahtar değeri ile kelimenin açıklaması verilmiştir.

```
<word form="able" wordnet_id="a-00510348"
  pos="JJ" sense="have the skills and qualifications to do things well"
  polarity="0.5" subjectivity="0.5" intensity="1.0" confidence="0.9" />
```

Şekil 3.14. Sentiment.xml dosyasından bir kayıt.

TextBlob kütüphanesi, bir kelimenin polarite, öznellik ve yoğunluk değerini hesaplarken XML dosyasında var olan kayıtlardan faydalanır. Birden fazla kelime olduğu zaman ise her kelimenin kendinden sonraki kelimeyi ne kadar etkilediğini hesaplamak için yoğunluk değerine bakmaktadır. Çizelge 3.4'te bir kaç örneğin polaritesi hesaplanmıştır.

Çizelge 3.4. Polarite hesaplama örnekleri

Cümle	Polarite	Nesnellik	Yoğunluk
Good	0.7	0.8	1.0
Very	0.2	0.3	1.3
Very good	0.9	0.8	

Çizelge 3.4'te gösterilen "very good" metninde polarite hesaplanırken "very" kelimesinin yoğunluk (kendinden sonraki kelimeyi etkileme derecesi) ile "good" kelimesinin polaritesi çarpılarak ($0{,}7 * 1{,}3 = 0{,}91$) hesaplanmıştır.

Metindeki negatif ekler ve kelimeler polariteyi "-0.5" ile çarpar (Çizelge 3.5).

Çizelge 3.5. Polaritenin hesaplanması

Cümle	Polarite	Nesnellik	Yoğunluk
Good	0.7	0.8	1.0
Not good	-0.35	0.6	

TextBlob kütüphanesi ayrıca bir metnin yazım yanlışlarını düzeltebilir (Bkz. Şekil 3.15), bir metni cümlelere ve kelimelere ayırabilir (Bkz. Şekil 3.16), tekil-çoğul dönüşümünü yapabilir(Bkz. Şekil 3.17), kelimeleri istenilen sayıda ayırabilir (Bkz. Şekil 3.18), kelimenin veya harfin frekansını (metinde kaç defa tekrarlandığını) hesaplayabilir

ve Google Translate API'yi kullanarak bir metnin hangi dilde yazıldığını ve çevirisini (Bkz. Şekil 3.19) yapabilir (Anonim, 2018b).

```
cumle = TextBlob("I havv goood speling!")
print(cumle.correct())
#I have good spelling!
```

Şekil 3.15. Metin düzeltici.

```
metin = TextBlob("Beautiful is better than ugly. "
                 "Explicit is better than implicit. "
                 "Simple is better than complex.")
>>> metin.words
WordList(['Beautiful', 'is', 'better', 'than', 'ugly',
    'Explicit', 'is', 'better', 'than', 'implicit',
    'Simple', 'is', 'better', 'than', 'complex'])
>>> zen.sentences
[Sentence("Beautiful is better than ugly."),
Sentence("Explicit is better than implicit."),
Sentence("Simple is better than complex.")]


for cumle in metin.sentences:
    print(cumle)
```

Şekil 3.16. Kelime ve cümlelere ayırma.

```
>>> hayvanlar = TextBlob("cat dog octopus")
>>> hayvanlar.words
WordList(['cat', 'dog', 'octopus'])
>>> hayvanlar.words.pluralize()
WordList(['cats', 'dogs', 'octopodes'])
```

Şekil 3.17. Tekil-Çoğul dönüşümü.

```
cumle = TextBlob("Now is better than never.")
cumle.ngrams(n=3)
#[WordList(['Now', 'is', 'better']),
# WordList(['is', 'better', 'than']),
# WordList(['better', 'than', 'never'])]
```

Şekil 3.18. Kelimeleri istenilen sayıda ayırma (N-Gram).

```
from textblob import TextBlob

metin = TextBlob("Bugün hava çok güzel")
translate(from_lang='TR-tr', to='en')
#The weather is very nice today
metin.detect_language()
#'TR-tr'
```

Şekil 3.19. Dil algılama ve çeviri.

Tweetler elde edilirken başka çalışmalarda kullanılabilme ihtimaline karşı *"tweet_id, username, created_at, tweet, cleaned_tweet, retweeted, user_location ve polarity"* bilgileri veri tabanına kaydedilmiştir. Ancak tez uygulamasında sadece "cleaned_tweet" ve "polarity" bilgileri kullanılmıştır. Üzerinde işlem yapılacak CSV dosyasının örnek ilk on kaydının gösterildiği Python kodu (Şekil 3.20) ve bu kodun ekran görüntüsü (Şekil 3.21) aşağıdaki gibidir.

```
import pandas as pd
csv ='cleen_tweet_btc.csv'
df = pd.read_csv(csv)
print df.head(10)
```

Şekil 3.20. CSV Dosyası okuma kodları.

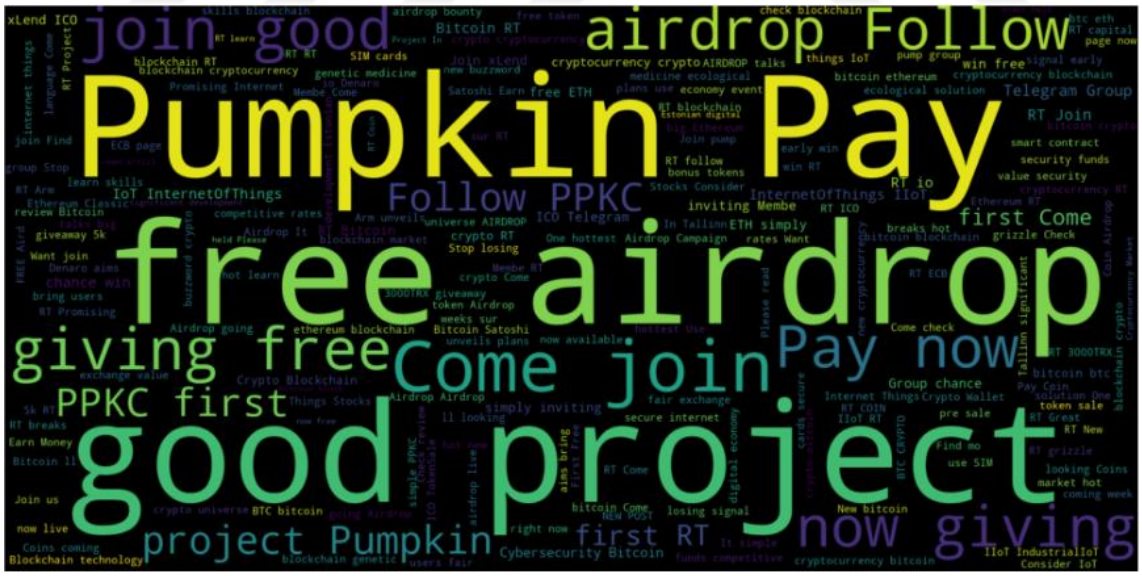
```
can@ubuntu:~/Downloads/python_file/yeni$ python ilkon.py
                                               tweet  target
0  Earn 0 00000005 BTC instantly 1 Retweet this m...       0
1  RT On Friday evening we went to the Crypto 010...       0
2  Buy BItcoin Iran just announced it bans use of...       0
3  RT Win 500 in CBC free just download the free ...       1
4  Are you a smart city initiative looking for io...       1
5  RT Petro Pre Sale 750 1 3000 127 17 announce 4...       0
6  RT plan Why Mine Bitcoin When You Can Mint Coi...       1
7  RT hur Here is our blog on How to add Hurify T...       0
8      IQeon Ethereum Benefits of Investing in IQeon       0
9  RT CBC Referral Program Get a 5 CashBet Coin t...       0
can@ubuntu:~/Downloads/python_file/yeni$
```

Şekil 3.21. CSV Dosyasının ekran görüntüsü.

Elde edilen tweetlerin temizlenmiş metni ve polaritesinin olduğu dosyadan kelime bulutu (Word Cloud) gösterimi için gerekli kod (Şekil 3.22) ve ekran çıktısı (Şekil 3.23) aşağıdaki gibidir. Şekil 21'de 7. satırda "1" yazılması durumunda olumlu, "0" yazılması durumunda nötr, "-1" yazılması durumunda ise olumsuz tweetlerin kelime bulutu elde edilir.

```
1  import pandas as pd
2
3
4  csv = 'clean_tweet_btc.csv'
5  df = pd.read_csv(csv)
6
7  neg_tweets = df[my_df.target==1]
8  neg_string = []
9  for t in neg_tweets.tweet:
10     neg_string.append(t)
11 neg_string = pd.Series(neg_string).str.cat(sep=' ')
12
13 from wordcloud import WordCloud
14
15 wordcloud = WordCloud(width=1600, height=800,max_font_size=200).generate(neg_string)
16 plt.figure(figsize=(12,10))
17 plt.imshow(wordcloud, interpolation="bilinear")
18 plt.axis("off")
19 plt.show()
```

Şekil 3.22. Tweetlerin kelime bulutu ile gösterimi.



Şekil 3.23. Kelime bulutu ekran görüntüsü.

En çok kullanılan para birimlerine ait elde edilen verilerin polarite karşılaştırması Çizelge 3.6'da gösterilmiştir.

Çizelge 3.6. Para birimlerinin karşılaştırılması

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

3.2.4. Tweetlerin n-Gram doğruluk skorları

N-gram, veriler üzerinde arama, sınıflandırma, karşılaştırma veya tekrar sayısını bulma gibi işlemler yaparken kelimelerin kaçar kaçar ayrıldığında en doğru skoru vereceğini hesaplamak için kullanılan bir yöntemdir. N-gram ifadesinde yer alan n değeri, tekrar derecesini ifade etmektedir. Konuyu bir örnekle açıklamak gerekirse, üzerinde işlem yapılacak cümlemiz "*Kartallar yüksek uçar ve yalnızdır*" cümlesinin n-gram a göre parçalanmış hali Çizelge 3.7'de gösterilmiştir.

Çizelge 3.7. N-Gram ile cümle parçalama

N Değeri	Ayrılmış kelimeler
1-Gram.Unigram	Kartallar, yüksek, uçar, ve, yalnızdır
2-Gram. Bigram	Kartallar yüksek, yüksek uçar, uçar ve, ve yalnızdır
3-Gram. Trigram	Kartallar yüksek uçar, yüksek uçar ve, uçar ve yalnızdır.
N-Gram. n-gram*	Kartallar yüksek uçar ve yalnızdır

*. 3 ten sonraki tüm parçalar n-gram terimi ile ifade edilir.

Tez uygulamasında n-gram doğruluk skorlarını karşılaştırmak üzere lojistik regresyon algoritması kullanılmıştır. Bu işlemlerin yapılabilmesi için, önceden CSV formatında kaydedilen tweetlerin sadece tweetlerin temizlenmiş metni ve polarite değerlerini alınarak Çizelge 3.8'deki gibi yeni bir CSV dosyaya kaydedilmiştir.

Çizelge 3.8. N-Gram doğruluk skoru verisi

tweet	target
Earn 0 00000005 BTC instantly 1 Retweet this message 2 Download and login 3 Check your gmail for bitcoin	0
Shares Drop on SEC Cryptocurrency Probe crypto altcoin ico	-1
Buy BItcoin Iran just announced it bans use of US dollar in trade	1
RT Win 500 in CBC free just download the free Pryze app Android or iOS wi	1
Are you a smart city initiative looking for iot sensor data Look no further We have a marketplace for you	0
RT Petro Pre Sale 750 1 3000 127 17 announce 40 USD 33 BTC 18 ETH 6 E	1
RT plan Why Mine Bitcoin When You Can Mint Coin With Less Impact on The Environment MintCoin is the eco friendly cryptocurr	0
RT hur Here is our blog on How to add Hurify Tokens to your MEW and MetaMask Wallet As Hurify Tokens are ERC20 compliant this	0
IQeon Ethereum Benefits of Investing in IQeon	0
RT CBC Referral Program Get a 5 CashBet Coin token reward on all purchases made by your friends through our Referral Progr	0
RT Games are on REGA says blockchain gives the solutions for the athletes too rega crowdsuranc	0
RT Join our Blockchain Asia energy renewables solarpower ico ethereum	1
Porsche are testing Blockchain technology Read more here gt Cryptonews BitcoinNews	1
Arbitrage opportunity has occurred The diff is 0 00001469 BTC 1 Buy GAME on poloniex 0 00016311 BT	-1

En iyi doğruluk oranlarını elde edebilmek için veriler her seferinde TF-IDF ile vektörlere dönüştürülürken 50000 adeti işlenmiş ardından unigram, bigram ve trigram ile doğruluk skorları ve bu skorları üretirken harcanan süre hesaplanmıştır. Şekil 3.24'te gösterilen kod parçasında temizlenmiş tweet metinleri "X" ve doğruluk skoru hesaplaması için polarite ise "Y" olarak etiketlenmiştir. Programda hataları gidermek için, temizleme sonrası elde edilen metinlerde hiç bir karakter yoksa boşluk (" ") ele alınmıştır. Verilerin %80'i eğitim ve %20'si test veri seti olarak hesaplanmıştır. Bu işlemlerin tümü durak kelimeler (stop words. cümleden çıkarılınca anlam kaybına uğratmayan kelimeler) dahil ve durak kelimeler hariç olmak üzere iki kez yapılmıştır.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

csv = 'clean_tweet_btc.csv'
my_df = pd.read_csv(csv)


x = my_df.tweet.fillna(' ')
y = my_df.target


from sklearn.cross_validation import train_test_split
x_train, x_validation_and_test, y_train, y_validation_and_test
= train_test_split(x, y, test_size=.02)
```

Şekil 3.24. Verilerin eğitim ve test ayrıştırılması.

Doğruluk skoru hesaplama işleminin birden fazla işlenmesinden dolayı bu işlem bir Şekil 3.25'te görülen şekilde bir fonksiyon yazılarak ihtiyaç duyulduğu kadar çalıştırılmıştır. Fonksiyon doğruluk skoru ve harcanan süreyi hesaplamaktadır. Fonksiyondaki "pipeline" değişkeni kelimeleri vektöre çevirirken kullanılan algoritma (TF-IDF) ve sınıflandırma algoritmasının (Lojistik Regresyon) bilgisini tutarak işlemleri o doğrultuda yapmaktadır.

```
def accuracy_summary(pipeline, x_train, y_train, x_test, y_test):
    t0 = time()
    sentiment_fit = pipeline.fit(x_train, y_train)
    y_pred = sentiment_fit.predict(x_test)
    train_test_time = time() - t0
    accuracy = accuracy_score(y_test, y_pred)
    print "accuracy score: {0:.2f}%".format(accuracy*100)
    print "train and test time: {0:.2f}s".format(train_test_time)
    return accuracy, train_test_time
```

Şekil 3.25. Doğruluk skoru ve süre fonksiyonu.

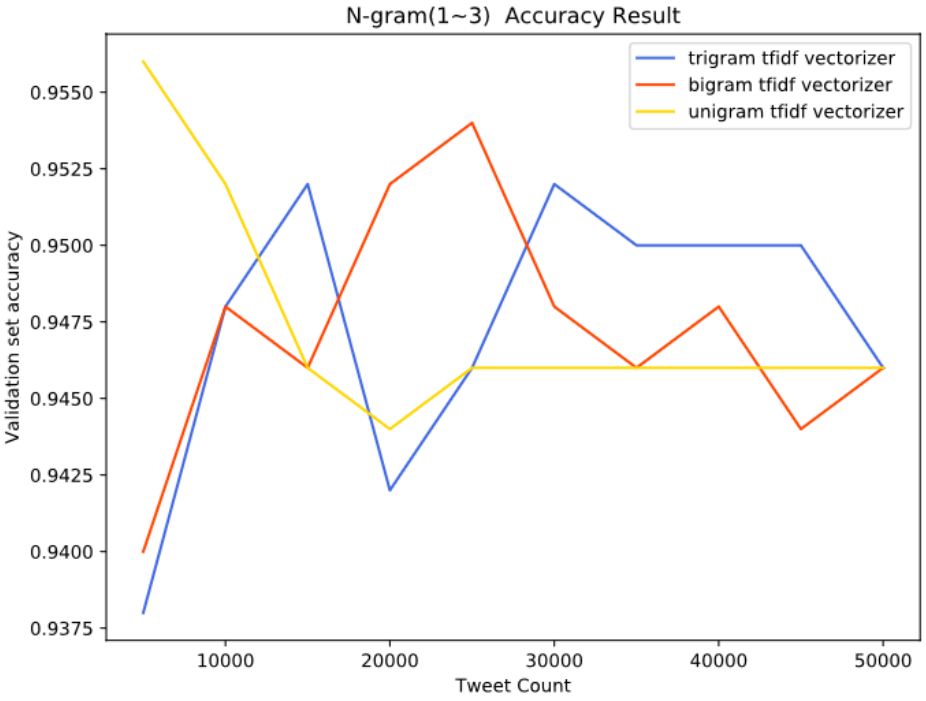
Bu işlem yapıldıktan sonra ulaşılan sonuçlar (Çizelge 3.9, Çizelge 3.10) ve bu sonuçların grafiği aşağıda gösterilmiştir (Bkz. Şekil 3.26, Şekil 3.27).

Çizelge 3.9. N-Gram doğruluk skoru ve harcanan süre (durak kelimeler ile, kripto para)

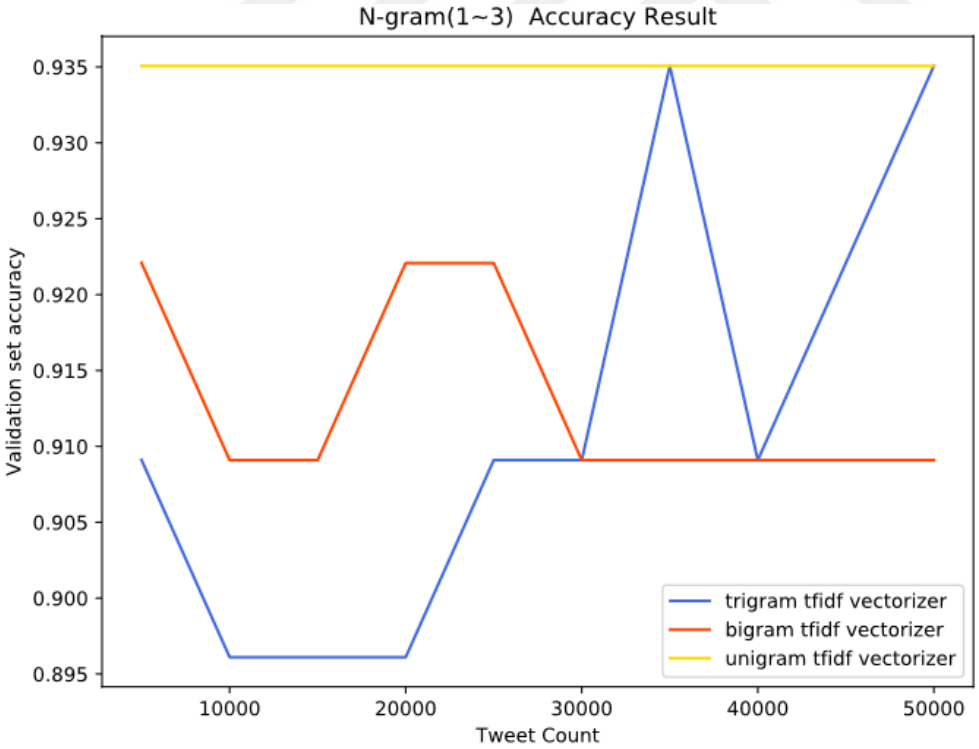
n-gram / Veri sayısı	Unigram		Bigram		Trigram	
	Doğruluk(%)	Süre(Sn)	Doğruluk(%)	Süre(Sn)	Doğruluk(%)	Süre(Sn)
1-5000	95.60	1.54	94.00	3.02	93.80	4.20
5000-10000	95.20	1.46	94.80	2.98	94.80	4.54
10000-15000	94.60	1.54	94.60	2.92	95.20	8.23
15000-20000	94.40	1.60	95.20	2.94	94.20	5.30
20000-25000	94.60	2.76	95.40	3.09	94.60	4.50
25000-30000	94.60	4.88	94.80	2.98	95.20	4.47
30000-35000	94.60	4.10	94.60	3.01	95.00	5.23
35000-40000	94.60	1.55	94.80	3.06	95.00	5.81
40000-45000	94.60	1.52	94.40	3.06	95.00	5.50
45000-50000	94.60	1.54	94.60	3.11	94.60	4.86
Ortalama	94.74	2.25	94.72	3.02	94.74	5.26

Çizelge 3.10. N-Gram doğruluk skoru ve harcanan süre (durak kelimeler ile, yapay zeka ve makine öğrenimi)

n-gram / Veri sayısı	Unigram		Bigram		Trigram	
	Doğruluk(%)	Süre(Sn)	Doğruluk(%)	Süre(Sn)	Doğruluk(%)	Süre(Sn)
1-5000	93,51	1,51	92,21	2,49	90,91	3,84
5000-10000	93,51	1,37	90,91	2,53	89,61	4,02
10000-15000	93,51	1,27	90,91	3,04	89,61	4,10
15000-20000	93,51	1,22	92,21	2,82	89,61	4,22
20000-25000	93,51	1,35	92,21	2,90	90,91	4,63
25000-30000	93,51	1,27	90,91	2,89	90,91	4,82
30000-35000	93,51	1,30	90,91	3,03	93,51	4,66
35000-40000	93,51	1,61	90,91	3,15	90,91	5,25
40000-45000	93,51	1,37	90,91	3,04	92,21	6,25
45000-50000	93,51	1,29	90,91	3,10	93,51	4,85
Ortalama	93,51	1,35	91,30	2,90	91,17	4,66



Şekil 3.26. N-Gram doğruluk skoru ve harcanan süre (durak kelimeler ile, kripto para).



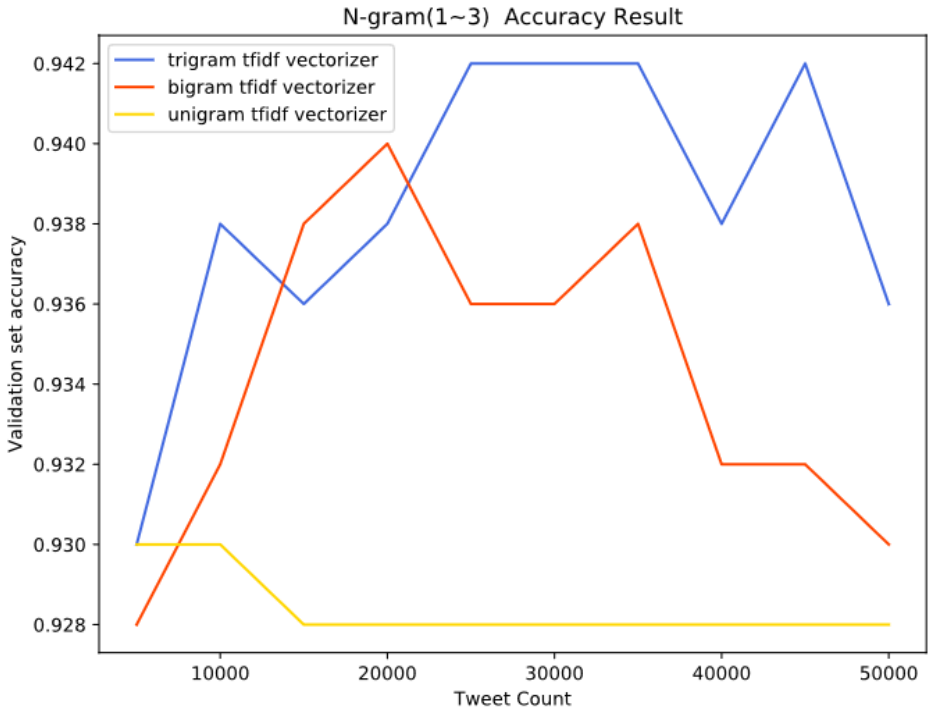
Şekil 3.27. N-Gram doğruluk skoru ve harcanan süre (durak kelimeler ile, yapay zeka ve makine öğrenimi).

Çizelge 3.11. N-Gram doğruluk skoru ve harcanan süre (durak kelimeler olmadan, kripto para)

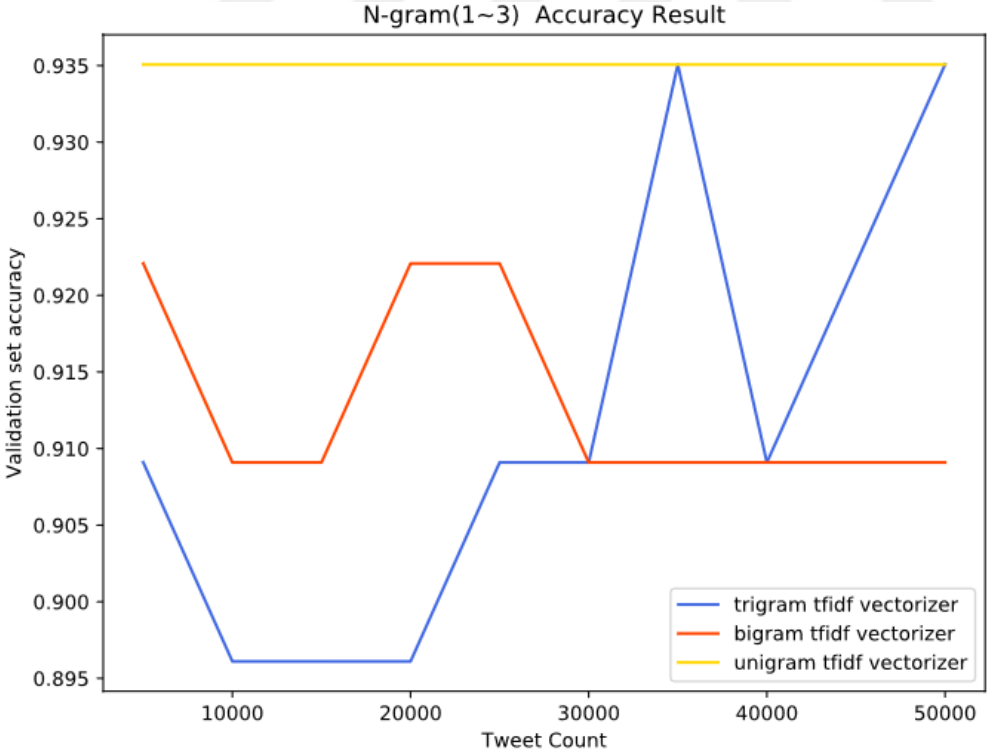
n-gram / Veri sayısı	Unigram		Bigram		Trigram	
	Doğruluk(%)	Süre(Sn)	Doğruluk(%)	Süre(Sn)	Doğruluk(%)	Süre(Sn)
1-5000	93.00	1.38	92.80	2.50	93.00	3.51
5000-10000	93.00	1.35	93.20	2.49	93.80	3.55
10000-15000	92.80	1.36	94.00	2.52	93.60	3.53
15000-20000	92.80	1.38	93.60	2.54	93.80	3.69
20000-25000	92.80	1.39	93.60	2.54	94.20	3.69
25000-30000	92.80	1.38	93.60	2.58	94.20	3.76
30000-35000	92.80	1.63	93.80	3.76	94.20	3.92
35000-40000	92.80	1.79	93.20	3.29	94.20	4.30
40000-45000	92.80	1.73	93.20	2.66	93.80	4.03
45000-50000	92.80	1.51	93.00	2.78	94.20	3.81
Ortalama	92.84	1.49	93.40	2.77	93.90	3.78

Çizelge 3.12. N-Gram doğruluk skoru ve harcanan süre (durak kelimeler olmadan, yapay zeka ve makine öğrenimi)

n-gram / Veri sayısı	Unigram		Bigram		Trigram	
	Doğruluk(%)	Süre(Sn)	Doğruluk(%)	Süre(Sn)	Doğruluk(%)	Süre(Sn)
1-5000	93,51	1,36	92,21	2,31	90,91	3,28
5000-10000	93,51	1,25	90,91	2,64	89,61	4,10
10000-15000	93,51	1,32	92,21	2,46	89,61	5,11
15000-20000	93,51	1,23	92,21	2,66	89,61	4,34
20000-25000	93,51	1,24	90,91	2,77	90,91	4,28
25000-30000	93,51	1,37	90,91	2,67	90,91	4,45
30000-35000	93,51	1,22	90,91	3,02	93,51	4,43
35000-40000	93,51	1,27	90,91	2,91	90,91	4,67
40000-45000	93,51	1,31	90,91	2,92	92,21	4,57
45000-50000	93,51	1,24	90,91	3,01	93,51	4,77
Ortalama	93,51	1,28	91,30	2,74	91,17	4,40



Şekil 3.28. N-Gram doğruluk skoru ve harcanan süre (durak kelimeler olmadan, kripto para).



Şekil 3.29. N-Gram doğruluk skoru ve harcanan süre (durak kelimeler olmadan, yapay zeka ve makine öğrenimi).

3.2.5. Verilerin sınıflandırılması ve kullanılan algoritmaların karşılaştırılması

Metin verileri, öngörülü modelleme için kullanılmaya başlamadan önce hazırlık gerektirmektedir. Elde edilen verilerden üzerinden çalışma yapılmayan alanlar çıkarıldıktan sonra sadece Tweet metni ile polaritenin olduğu bir CSV dosyasına kaydedilmektedir.

Metin değerlerin özellik çıkarma (veya vektörlendirme) olarak adlandırılan bir makine öğrenme algoritmasına girdi olarak kullanması için tamsayılar veya kayan nokta değerleri olarak kodlanması gerekir.

Metinleri vektörlendirme için Scikit-Learn kütüphanesinden faydalanılmıştır. Bu kütüphanede vektörlendirme için CountVectorizer (kelime sayma yöntemi) ve TfIdfVectorizer (kelime frekansı ve ters logaritma yöntemi) yöntemleri ile belge vertörize edildikten sonra %80'i eğitim ve %20'si test veri seti olmak üzere ikiye ayrılmaktadır(Bkz. Şekil 3.24). Daha sonra en çok kullanılan sınıflandırma algoritmalarına (*"Logistic Regression", "Linear SVC", "Multinomial NB", "Bernoulli NB", "Ridge Classifier", "Passive Aggresive Classifier"*) tabi tutulmaktadır (Şekil 3.30). Daha sonra bu sınıflandırmaların süre ve doğruluk skorları karşılaştırılmaktadır (Şekil 3.30).

```
from sklearn.svm import LinearSVC
from sklearn.naive_bayes import MultinomialNB, BernoulliNB
from sklearn.linear_model import RidgeClassifier
from sklearn.linear_model import PassiveAggressiveClassifier


from sklearn.feature_selection import SelectFromModel

names = ["Logistic Regression", "Linear SVC", ,"Multinomial NB",
         "Bernoulli NB", "Ridge Classifier",  "Passive-Aggresive"]
classifiers = [
    LogisticRegression(),
    LinearSVC(),
    MultinomialNB(),
    BernoulliNB(),
    RidgeClassifier(),
    PassiveAggressiveClassifier()

    ]
```

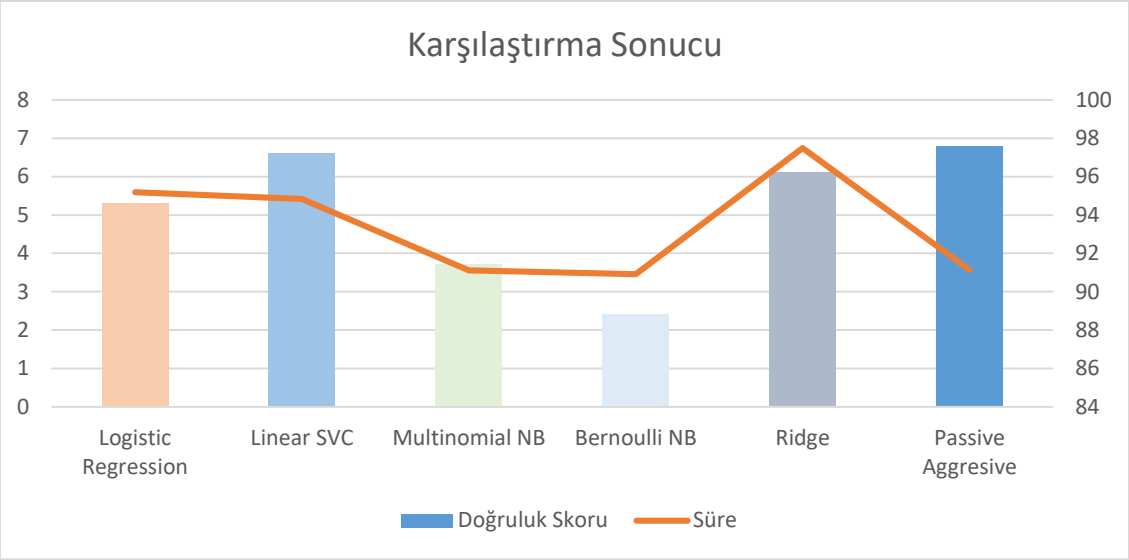
Şekil 3.30. Sınıflandırma algoritmalarının kullanımı.

```
tvec = TfidfVectorizer()
def classifier_comparator(vectorizer=tvec, n_features=10000,
                          stop_words=None, ngram_range=(1, 1),
                          classifier=zipped_clf):
    result = []
    vectorizer.set_params(stop_words=stop_words, max_features=n_features,
                          ngram_range=ngram_range)
    for n,c in classifier:
        checker_pipeline = Pipeline([
            ('vectorizer', vectorizer),
            ('classifier', c)
        ])
        print "Validation result for {}".format(n)
        print c
        clf_accuracy,tt_time =
        accuracy_summary(checker_pipeline,
                         x_train, y_train, x_validation, y_validation)
        result.append((n,clf_accuracy,tt_time))
    return result
trigram_result = classifier_comparator(n_features=100000,ngram_range=(1,3))
```

Şekil 3.31. Sınıflandırma algoritmalarının karşılaştırılması.

Çizelge 3.13. Sınıflandırma algoritmaları karşılaştırma sonuçları (kripto para)

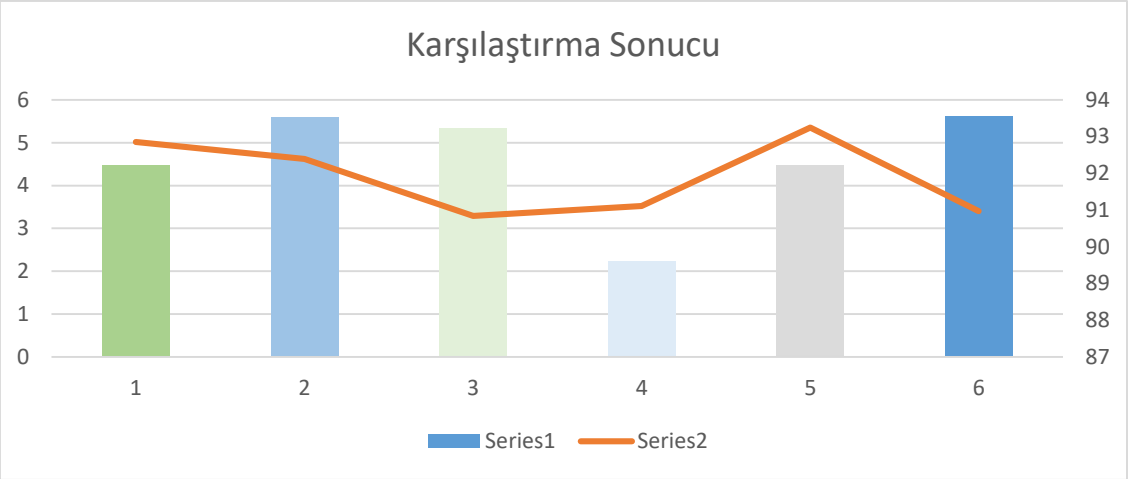
Sınıflandırma Algoritması	Doğruluk Skoru(%)	Süre(Saniye)
Logistic Regression	94,60	5,59
Linear SVC	97,20	5,42
Multinomial NB	91,40	3,56
Bernoulli NB	88,80	3,46
Ridge	96,20	6,74
Passive Aggresive	97,60	3,57



Şekil 3.32. Sınıflandırma algoritmaları karşılaştırma sonuçları (kripto para).

Çizelge 3.14. Sınıflandırma algoritmaları karşılaştırma sonuçları (yapay zeka ve makine öğrenimi)

Sınıflandırma Algoritması	Doğruluk Skoru(%)	Süre(Saniye)
Logistic Regression	92,21	5,02
Linear SVC	93,51	4,62
Multinomial NB	93,21	3,29
Bernoulli NB	89,61	3,52
Ridge	92,21	5,35
Passive Aggresive	93,55	3,40



Şekil 3.33. Sınıflandırma algoritmaları karşılaştırma sonuçları (yapay zeka ve makine öğrenimi).

4. BULGULAR

Çizelge 3.9 & Çizelge 3.10 ve Çizelge 3.11 & Çizelge 3.12'den anlaşılacağı üzere doğruluk skoru anlamında unigram, bigram ve trigram arasında kayda değer bir fark yoktur. Ancak bigram ve trigram'da kelimelerin birden fazla işlenmesinden dolayı harcanan süre daha fazladır. Farklı veri aralıklarındaki farklı doğruluk skoru ve harcanan süredeki farklılıklar ise, o aralığa denk gelen verilerin uzunluğu ile ilgili olup önemli bir değer farkına sahip değildir.

Toplamda 250.000 adet kripto para kategorisindeki metin değerler (tweet) için elde edilen ortalama doğruluk skorları ve harcanan süreler Çizelge 4.1, Çizelge 4.2 ve Çizelge 4.3'te gösterilmiştir. Unigram, bigram ve trigram doğruluk skorları (%94.74, %94.72, %94.74) arasında ciddi bir yakınlık mevcuttur. Ancak süreler anlamında (11.25 sn, 15.10 sn, 26.30 sn) ciddi bir fark (trigram, unigram süresinden iki kat fazla) gözlemlenmiştir. Aynı algoritmalar, makine öğrenimi ve yapay zeka kategorilerine ait 50.000 veriye uygulandığında doğruluk skorlarının ve sürelerinin oransal olarak çok yakın olduğu gözlemlenmiştir. Makine öğrenimi ve yapay zeka kategorisindeki unigram, bigram ve trigram doğruluk skorları (%95.51, %93.30, %91.17) bir birine yakın olmasına rağmen unigram daha iyi bir sonuç üretmiştir. Bununla beraber bu işlem sırasında harcanan sürelerde (unigram -1.35 sn., bigram-2.90 sn., trigram-4.66 sn.) ciddi farklılıklar gözlemlenmiştir.

Durak kelimeler (stopwords) olarak bilinen ve metinlerden çıkarılınca olumlu/olumsuz anlamda metne etki etmeyen kelimeler (ben, kadar, için vb.) çıkarılınca algoritma sürelerinde ciddi bir azalma gözlemlenmiştir. Özellikle kripto para kategorisinde unigram ile yapılan sınıflandırma da neredeyse yarı yarıya (11,25sn → 7,45sn) süre düşümü gözlemlenmiştir. Buna karşın doğruluk skorunda küçük bir iyileşme olmuştur. Ancak sürede oluşan farkın yanında küçük kalmıştır.

Çizelge 4.1. Unigram skorları

Sınıflandırma & Unigram	Unigram Değerleri			
	D'li*		D'siz**	
	Skor	Süre	Skor	Süre
Kripto Para	94,74	11,25	92,84	7,45
MÖ ve YZ***	93,51	1,35	93,51	1,28

* . Durak Kelimeler ile
** . Durak Kelimeler Olmadan
*** . Makine Öğrenmesi ve Yapay Zeka

Çizelge 4.2. Bigram Skorları

Sınıflandırma & Bigram	Bigram Değerleri			
	D'li		D'siz	
	Skor	Süre	Skor	Süre
Kripto Para	94,72	15,10	93,40	13,85
MÖ ve YZ	93,30	2,90	91,30	2,74

Çizelge 4.3. Trigram skorları

Sınıflandırma & Trigram	Trigram Değerleri			
	D'li		D'siz	
	Skor	Süre	Skor	Süre
Kripto Para	94,74	26,30	93,90	18,90
MÖ ve YZ	91,17	4,66	91,17	4,40

Bütün parametreler dikkate alındığında en ideal algoritma durak kelimeler çıkarıldıktan sonra unigram olarak görülmüştür.

5. TARTIŞMA VE SONUÇ

Günümüzde internetin yaygınlaşması ile birlikte büyük veride ciddi artışlar olmuştur. Her iki senede, kendinden önceki bütün bilgiler kadar bilgi üretilmektedir. Özellikle Web 2.0 ile beraber internete bağlanan bütün kullanıcılar veri üretimine katılmıştır. Patlama şeklinde artan verilerden anlamlı bilgiye ulaşmak artık zorunlu hale gelmiştir. Sosyal medya bilgileri birçok olaya yön vermeye başlamış ve şirketler, eğitim kurumları veya kişiler, sosyal medyadaki kullanıcı eğilimlerine göre kendilerine yön belirlemektedir.

Bu tez uygulamasında, hem geniş kitleler tarafından kullanılan Twitter verileri üzerinde çalışılmış hem de son yıllarda popülaritesini arttıran kripto para, makine öğrenmesi ve yapay zeka konularına ait veriler işlenmiştir. Belirlenen konularda yazılan toplam 350 bin tweet için yaklaşım durumları olumlu/olumsuz/tarafsız olarak sınıflandırılmış, makine öğrenmesinde kullanılan metin sınıflandırma algoritmalarına tabii tutularak doğruluk skorları elde edilmiş ve bu işlemler esnasında bilgisayarın harcadığı süre hesaplanmıştır. Metin sınıflandırma yapılırken unigram, bigram ve trigram yöntemleri ile analiz edilmiş, trigram doğruluk olarak çok az (yaklaşık %1-2) daha iyi bir doğruluk skoruna sahip olmasında rağmen süre bazında kötü sonuç elde edilmiştir. Trigram'da çok az da olsa daha iyi doğruluk elde edilmesi, kelimelerin anlamlarını, önündeki ve sonrasındaki kelimelerin etkilenmesinden kaynaklanmaktadır. Ancak yukarıda da belirtildiği gibi bu ciddi bir fark değildir. Trigram'ın süre anlamında kötü sonuç üretmesi ise, baştaki ve sondaki iki kelime hariç diğer kelimelerin üçer kez analiz edilmesinden kaynaklanmaktadır. Bu sebeple trigram süre konusunda unigram'ın çok gerisinde kalmaktadır.

Metin sınıflandırma algoritmalarının kullanımında en iyi sürelere Naif Bayes teoremini temel alan Multinomial ve Pasif-Agresif yöntemleri olsa da harcanan sürede Pasif-Agresif algoritmasının daha iyi bir sonuç ürettiği gözlemlenmiştir. Pasif-Agresif algoritmasının daha iyi sürede skor elde etmesinin nedeni, yeni ele aldığı verinin özelliklerine baktığında bir önceki veri ile aynı özelliklere sahip ise algoritmaya tabii tutmadan bir önceki veri ile aynı sınıfa dahil etmesinden kaynaklanmaktadır.

Verinin içeriğine ve uzunluğuna göre farklı (ama yakın) sonuçlar elde edilse de metin ayrıştırma için en iyi sonucun unigram ve metin sınıflandırmada ise Pasif-Agresif yönteminin en iyi sonuca vardığı gözlemlenmiştir.

Sonuç olarak, tez kapsamında elde edilen bulgular neticesinde sosyal medya platformlarından elde edilen veriler üzerinde anlamlı bilgilere ulaşmanın mümkün olacağı değerlendirilmektedir. İleriki çalışmalarda güncel konular üzerinde sosyal medya kullanıcılarının duyarlılık analizi derin öğrenme uygulamarıyla tespit edilmeye çalışılacaktır.

KAYNAKLAR

Alikılıç, Ö., A., 2011. Halkla ilişkiler 2,0. Sosyal Medyada Yeni Paydaşlar, Yeni Teknikler. Efil Yayınevi.

Alper, A., 2012. Sosyal Ağlar, Pelikan Yayıncılık, Ankara. 83.

Anonim, 2013. https.//idc-cema,com/eng/events/50533-idc-big-data-and-business-analytics-forum-2013?g_clang=TR. Erişim Tarihi. 10.12.2017.

Anonim, 2014. http.//www.linearsvm.com/. Erişim Tarihi. 14.06.2017.

Anonim, 2014. https.//en.wikipedia.org/wiki/Support_vector_machine. Erişim Tarihi. 14.06.2017.

Anonim, 2017a. http.//whatsabyte.com. Erişim Tarihi. 14.12.2017.

Anonim, 2017b. https.//tr.wikipedia.org/wiki/Makine_öğrenimi. Erişim Tarihi. 15.01.2018.

Anonim, 2017c. Sentiment Analyses. https.//monkeylearn.com/sentiment-analysis/. Erişim Tarihi. 15.03.2018.

Anonim, 2017d. https.//tr.wikibooks.org/wiki/Linux_İşletim_Sistemi/Linux_Nedir%3F. Erişim Tarihi. 15.03.2018.

Anonim, 2017e. https.//wiki.ubuntu-tr.net/index.php?title=Ubuntu_nedir%3F. Erişim Tarihi. 15.03.2018.

Anonim, 2017f. https.//developer.twitter.com/en/docs/api-reference-index. Erişim Tarihi. 15.03.2018.

Anonim, 2017g. https.//www.python.org/about/apps/. Erişim Tarihi. 15.03.2018.

Anonim, 2018a. https.//www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Erişim Tarihi. 15.09.2018.

Anonim, 2018b. https.//textblob.readthedocs.io/en/dev/quickstart.html. Erişim Tarihi. 15.03.2018.

Anonim, 2018c. https.//www.bonaccorso.eu/2017/10/06/ml-algorithms-addendum-passive-aggressive-algorithms/. Erişim Tarihi. 15.03.2018.

Atasoy, D., 2001. "***Lojistik Regresyon Analizinin İncelenmesi Ve Bir Uygulaması"*** **(**Yüksek Lisans Tezi). Cumhuriyet Üniversitesi. Sosyal Bilimler Enstitüsü. Sivas.

Aydın, S., 2007. ***Veri Madenciliği Ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama*** (Doktora tezi), Anadolu Üniversitesi, Eskişehir.

Baykal, A., 2006. "Veri Madenciliği Uygulama Alanları", D.Ü. Ziya Gökalp Eğitim Fakültesi Dergisi, (7). 95-107.

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. 21.12.2017.

Bifet, A., Frank, E., 2010. Sentiment Knowledge Discovery İn Twitter Streaming Data, In International Conference On Discovery Science , Springer, Berlin, Heidelberg. 1-15.

Bisk, 2017. "What is Big Data?", https.//www.villanovau.com/resources/bi/what-is-big-data. Erişim Tarihi. 20.12.2017.

Cavallo, A., 2017. ***"Scraped Data and Sticky Prices. Frequency, Hazards, and Synchronization"*** (Doktora Tezi). Harvard University. https.//www.gartner.com/doc/2626815/predicts--big-data. Erişim Tarihi. 21.12.2017.

Ceci, M., 2005. Naive Bayesian Learning From Structural Data, Ph, D, dissertation, Dipartimento di Informatica, University of Bari. Italy.

Cortes, C., Vapnik, V., 1995. Support-vector Network, Machine Learning, 20.273–297.

Çelik, F., M., 2018. Bir Nereden Nereye Hikâyesi. Sosyal Medya. https.//brandingturkiye.com/bir-nereden-nereye-hikayesi-sosyal-medya/. Erişim Tarihi. 25.03.2018.

Davenport, T., 2012. HBR. https.//hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/. Erişim Tarihi. 10.01.2018.

David Zimbra, M., Lee, G., 2016. "Brand-Related Twitter Sentiment Analysis using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks". IEEE 1530-1605, 2016.

Demirbas, M., Akcora, C., G., 2010. Twitter. Roots, Influence, Applications, http.//cse.buffalo.edu/tech-reports/2010-03.pdf. Erişim Tarihi. 15.10.2018.

Dirin, B., 2014. "Big Data Nedir?", https.//netvent,com/big-data-nedir/. Erişim Tarihi. 19.12.2017.

Domenico, T., 2013. "Clouds for Scalable Big Data Analytics", In. *IEEE Computer Society*. (46). 98–101

Ege, B., 2011. Yeni Bilgi Modelleme Ve Programlama Felsefesiyle Semantik Web, Bilim Ve Teknik Dergisi, Aralık Sayısı. 36-39.

Getting, B., 2017. https.//practicalecommerce,com/Basic-Definitions-Web-1-0-Web-2-0-Web-3-0. Erişim Tarihi. 25.03.2018.

Hayes, A., S., 2017. Cryptocurrency Value Formation. An Empirical Study Leading To A Cost Of Production Model For Valuing Bitcoin, Telematics And Informatics, 34(7). 1308-1321.

İnanç, V., 2018. "Sosyal Medya Nedir ? Sosyal Medya Siteleri (Sosyal Ağlar) Nelerdir ?". https.//www.mediaclick.com.tr/blog/sosyal-medya-nedir-sosyal-medya-siteleri-sosyal-aglar-nelerdir. Erişim Tarihi. 28.03.2018.

Halima Banu, S., Chitrakala, S., 2016. "Trending Topic Analysis Using Novel Sub Topic Detection Model", (IEEE) ISBN- 978-1-4673-9745-2.

Hvistendahl, M., 2017. "Can 'Predictive Policing' Prevent Crime Before İt Happens?", http.//www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens. Erişim Tarihi. 17.12.2017.

James, J., 2017. "Data Never Sleeps", https.//www.domo.com/blog/data-never-sleeps-4-0/. Erişim Tarihi. 15.11.2017.

Karabulut, N., 2009. Yeni Medya Teknolojileri Ve Halkla İlişkiler, İstanbul, Beykoz Lojistik, Meslek Okulu Yayınları.

Kibriya, A., M., Frank, E., Pfahringer, B., Holmes, G., 2004. Multinomial Naive Bayes For Text Categorization Revisited, In *Australasian Joint Conference on Artificial Intelligence* , Springer, Berlin, Heidelberg. 488-499.

Kwak, H., Lee, C., Park, H., Moon, S., 2010. "What is Twitter, a Social Network or a News Media?,". Proceedings of the 19th International Conference on World Wide Web, New York, NY, USA. 591–600.

Karcıoğlu, F., Kurt, E., 2013. "Örgütsel İletişimin Etkinliği Açısından Kurumsal Bloglar ve Birkaç Kurumsal Blogun İncelenmesi", Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, (23). 3.

Laney, D., 2001. "*3D Data Management. Controlling Data Volume, Velocity and Variety*", Tech, rep, Meta Group,

Linch, M., 1998. https.//en.wikipedia.org/wiki/Unstructured_data. Erişim Tarihi. 19.12.2017.

Mamgain, N., Ekta, M., 2016. Ankush Mittal and Gaurav Bhatt, "Sentiment Analysis of Top Colleges in India Using Twitter Data", (IEEE) ISBN -978-1-5090-0082-1.

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

McCallum, A., & Nigam, K, (1998, July), A Comparison Of Event Models For Naive Bayes Text Classification, In *AAAI-98 Workshop On Learning For Text Categorization.* Vol, 752, (1). 41-48.

Pang, B., Lee, L., 2008. Opinion Mining And Sentiment Analysis, *Foundations and Trends® in Information Retrieval*, *2*(1–2). 1-135.

Peter, I., 2010. http.//www.nethistory.info/History%20of%20the%20Internet/index.html, Erişim Tarihi. 01.04.2018.

Pileggi, S. F., Fernandez-Llatas, C., Traver, V. 2012. When The Social Meets The Semantic. Social Semantic Web Or Web 2.5.

Poynter, R., 2012. *İnternet ve Sosyal Medya Araştırmaları El Kitabı*, İstanbul, Yayınevi; Optimist. 30.

Ramos, J., 2003. Using Tf-İdf To Determine Word Relevance İn Document Queries, In *Proceedings Of The First İnstructional Conference On Machine Learning*, (242). 133-142.

Schneider, C., 2016. "The Biggest Data Challenges That You Might Not Even Know You Have", https.//www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/. Erişim Tarihi. 15.03.2018.

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,

Sahayak, V., 2015. Vijaya Shete and Apashabi Pathan, "Sentiment Analysis on Twitter Data", (IJIRAE) ISSN. 2349-2163.

Sarno, D., 2009. Jack Dorsey On The Twitter Ecosystem, Journalism And How To Reduce Reply Spam, Part II, Los Angeles Times, 19.

Savaş, S., Topaloğlu, N., Yılmaz, M., 2012. Veri Madenciliği Ve Türkiye'deki Uygulama Örnekleri.

Scott, J., 2005. *BBS. The Documentary*. http.//www.wikizeroo.net/wiki/en/Internet_Archive. Erişim Tarihi. 15.03.2018.

Swan, M., 2015. Blockchain. Blueprint For A New Economy, O'Reilly Media, Inc.

Utkun, G., 2012. MICROSOFT Blog. "Büyük Veri Nedir". http.//blog.microsoft.com.tr/?p=4068. Erişim Tarihi. 15.02.2017.

Wikibooks, 2017. https.//tr.wikibooks.org/wiki/Linux_İşletim_Sistemi/Linux_Nedir%3F. Erişim Tarihi. 10.02.2018.

Vinod, H., D., Ullah, A., 1981. "Recent Advances in Regression Methods", Marcel Dekker Inc. New York. 30-45.

Yaman, A., 2016. "Python Dili Nedir?". https.//www.python.tc/python-nedir/. Erişim Tarihi. 16.05.2018.

Yu, H., F., Huang, F., L., Lin, C., J., 2011. Dual Coordinate Descent Methods For Logistic Regression And Maximum Entropy Models, *Machine Learning*, *85* (1-2). 41-75.

ÖZGEÇMİŞ

1984 yılında Muş'ta doğdu. Orta öğrenimini Van'da tamamladı. 2004 yılında Mersin Üniversitesi Bilgisayar Mühendisliği Bölümü'nden mezun oldu. 2007 yılından beri Van Yüzüncü Yıl Üniversitesi'nde Öğretim Görevlisi olarak çalışmaya devam etmektedir.

T.C
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
LİSANSÜSTÜ TEZ ORİJİNALLİK RAPORU

Tarih: 20/12/2018

Tez Başlığı / Konusu:

BÜYÜK VERİ ARAÇLARI KULLANARAK SOSYAL MEDYADA HİS ANALİZİ YAPMA

Yukarıda başlığı/konusu belirlenen tez çalışmamın Kapak sayfası, Giriş, Ana bölümler ve Sonuç bölümlerinden oluşan toplam 5 sayfalık kısmına ilişkin, 20/12/2018 tarihinde şahsım/tez danışmanım tarafından **turnitin** intihal tespit programından aşağıda belirtilen filtreleme uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 4 (dört) dır.

Uygulanan filtreler aşağıda verilmiştir:
- Kabul ve onay sayfası hariç,
- Teşekkür hariç,
- İçindekiler hariç,
- Simge ve kısaltmalar hariç,
- Gereç ve yöntemler hariç,
- Kaynakça hariç,
- Alıntılar hariç,
- Tezden çıkan yayınlar hariç,
- 7 kelimeden daha az örtüşme içeren metin kısımları hariç (Limit inatch size to 7 words)

Van Yüzüncü Yıl Üniversitesi Lisansüstü Tez Orijinallik Raporu Alınması ve Kullanılmasına İlişkin Yönergeyi inceledim ve bu yönergede belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini bilgilerinize arz ederim.

20.12.2018

Adı Soyadı: Mehmet Can ERDOĞAN

Öğrenci No:.....

Anabilim Dalı: Elektrik Elektronik ABD

Programı: .....

Statüsü: Y. Lisans ☒ Doktora ☐

DANIŞMAN ONAYI
UYGUNDUR

Murat CANAYAZ
Dr. Öğr. Üyesi

ENSTİTÜ ONAYI
UYGUNDUR

Prof.Dr.Suat ŞENSOY
Enstitü Müdürü