T.C.
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANABİLİM DALI
SAYISAL YÖNTEMLER BİLİM DALI

DOKTORA TEZİ

BÜYÜK VERİ VE İŞLETME ANALİTİĞİ: SOSYAL MEDYA VE DUYGU ANALİZİ İLE BİR ÖNGÖRÜ MODELİ

BURCU KARAÖZ
2502140266

Tez Danışmanı
PROF. DR. UMMAN TUĞBA ŞİMŞEK GÜRSOY

İSTANBUL, 2018

T.C.
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ

DOKTORA
TEZ ONAYI

ÖĞRENCİNİN;

Adı ve Soyadı	:	BURCU KARAÖZ	**Numarası**	:	2502140266
Anabilim Dalı / Anasanat Dalı / Programı	:	SAYISAL YÖNTEMLER	**Danışmanı**	:	DOÇ.DR.UMMAN TUĞBA ŞİMŞEK GÜRSOY
Tez Savunma Tarihi	:	02-03-2018	**Saati**	:	11.00
Tez Başlığı	:	BÜYÜK VERİ VE İŞLETME ANALİTİĞİ: SOSYAL MEDYA VE DUYGU ANALİZİ İLE BİR ÖNGÖRÜ MODELİ			

TEZ SAVUNMA SINAVI, İÜ Lisansüstü Eğitim-Öğretim Yönetmeliği'nin **50. Maddesi** uyarınca yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin **KABULÜNE** OYBİRLİĞİ / ~~OYÇOKLUĞUYLA~~ karar verilmiştir.

JÜRİ ÜYESİ	İMZA	KANAATİ (KABUL / RED / DÜZELTME)
PROF.DR. MEHPARE TİMOR		Kabul
PROF.DR. AHMET METE ÇİLİNGİRTÜRK		Kabul
DOÇ.DR.UMMAN TUĞBA ŞİMŞEK GÜRSOY	Tuğba	Kabul
DOÇ.DR.SONA MARDİKYAN	S Mardik	Kabul
DOÇ.DR.TARIK KÜÇÜKDENİZ		Kabul

YEDEK JÜRİ ÜYESİ	İMZA	KANAATİ (KABUL / RED / DÜZELTME)
PROF.DR.ZUHAL TANRIKULU		
DOÇ.DR.SEDA TOLUN TAYALI		

Versiyon: 1.0.0.2-61559050-302.14.06

ÖZ

BÜYÜK VERİ ve İŞLETME ANALİTİĞİ: SOSYAL MEDYA ve DUYGU ANALİZİ ile BİR ÖNGÖRÜ MODELİ

Burcu Karaöz

İşletme analitiği ile işletmelerin içyapısında biriken veya çeşitli dış kaynaklardan toplanan verilerin derlenmesi, depolanması, düzenlenmesi ve analiz edilmesiyle anlamlı bilgiler ortaya çıkararak, işletmeler için fayda yaratabilmektedir. Gelişen teknoloji, önemli dış kaynaklardan sosyal medya kanalların aracılığıyla anket ya da diğer görüşme yöntemleri kullanılarak sayıca ulaşılması mümkün olamayacak kadar kişisel görüşe çok kısa sürelerde ulaşabilmeyi ve duygu analizi teknikleri aracılığıyla bu görüşleri yapısal formata çevirebilmeyi mümkün kılmaktadır. Bu doğrultuda, çalışmada bir işletme ve ürünleri hakkında sosyal medyadan toplanan verilerin, iletiyi yazan kişilerin ilgili işletme ve ürünleri hakkında belirttikleri görüşlerin olumlu, olumsuz ve nötr olarak nitelendirilebilecek görüşlerden hangisine dahil olduğunun araştırılabilmesi için duygu analizi algoritması geliştirilmiştir. Sosyal medyadan edinilen verilerin nitelik ve nicelik olarak işletmenin gerçek değerleri ile ilişkili olup olmadığının incelenmesi için kümeleme analizi uygulanmıştır. Daha sonra her ürün için derlenen ileti sayısı, iletilerin duyguları ve ürünlerin kendine has diğer özellikleri ile birlikte tahmin modelinde girdi olarak kullanılmıştır. Böylece işletmeye ilgili ürünler hakkında karar alma ve yeni ürün piyasaya sürme aşamalarında ya da geliştirecekleri pazarlama stratejilerinde faydalanabilecekleri bir sistem kurulması hedeflenmiştir. Bu amaçla karar ağacı, destek vektör makineleri ve yapay sinir ağları yöntemleri kullanılmıştır ve tahmin performansları incelendiğinde, her üç yöntemin de anlamlı sonuçlara ulaşılmasını sağladığı anlaşılmıştır. Bu üç yöntem karşılaştırıldığında ise yapay sinir ağlarının en yüksek tahmin performansına sahip olduğu görülmüştür. Böylece sosyal medyadan elde edilen verilerin nitelik ve nicelik olarak değerlendirilmesinin işletmelere önemli etkiler yaratabileceği, mevcut durum analizi ya da geleceğe yönelim amacıyla kullanılabileceği ortaya çıkarılmıştır.

Anahtar Kelimeler: büyük veri, işletme analitiği, sosyal medya analitiği, duygu analizi

ABSTRACT

BIG DATA and BUSINESS ANALYTICS: A FORECASTING MODEL via SOCIAL MEDIA and SENTIMENT ANALYSIS

Burcu Karaöz

Business analytics creates value and gains insight into knowledge via collecting, warehousing, organizing and analyzing the data that obtained from operations or external resources. Social media, one of the valuable external resources, provides large datasets that is inaccessible to collect with survey or other traditional techniques. Aim of this study is collecting tweets -which are about a company and its products/services-, construe the opinions of tweets and revealing beneficial information for the company. For this purpose, we chose a TV channel as a company and generated a forecasting model to estimate rating values of programs. Beginning, tweets are gathered about the channel and the programs and attitude of tweets are determined via proposed sentiment analysis algorithm. Proposed algorithm assumes that each dataset needs specific opinion lexicons, so semi-supervised sentiment analysis algorithm is generated. The results of sentiment analysis and features of programs such as audiences, genre, airing time and star value are used to forecast the rating of next broadcast. In our model, the forecasting problem is converted into a classification problem-rather than forecasting the point estimate of ratings- ratings in one of five categories is classified. The model designed to predict the expected rating range of a program for the next broadcast. Comparison of neural networks, support vector machines and decision tree algorithms for this model shows that all algorithms gives significant results and neural network algorithm has the best performance. Results of this study indicate that social media makes accessible to consumer sentiment and analysizng patterns in this data ensures insightful decision making for the companies.

Keywords: big data, business analytics, social media analytics, sentiment analysis

ÖNSÖZ

Bana güvenip her zaman destek olan danışmanım Tuğba Gürsoy'a,

Bana her konuda destek olan hocam Şevkinaz Gümüşoğlu'na,

Bana zaman ayırıp tezimde önümü açan hocam Güzin Özdağoğlu'na,

Üzerimde emeği geçen tüm hocalarıma,

Aileme,

Alp, Burcu, Yeliz, Ece, Caner, Aylin'e,

Göktürk'e sonsuz teşekkürlerimi sunarım.

Burcu Karaöz

İstanbul, 2018

İÇİNDEKİLER

TABLO LİSTESİ

ŞEKİL LİSTESİ

KISALTMALAR LİSTESİ

20+ABC1 : 20 Yaş Üzeri Ve Sosyoekonomik Statü Grubunda A, B Ve C1'e Dahil Olan Bireyler

5+ : 5 Yaş Üzeri Bireyler

API : Application Programming Interface

CART : Classification And Regression Trees

CHAID : Chi-Square Automatic Interaction Detection

DIKW : Data, Information, Knowledge, Wisdom

DJIA : Dow Jones Borsası Endüstri Endeksi

DVM : Destek Vektör Makineleri

FN : False Negative

FP : False Positive

GOT : Get Old Tweets

HDFS : Hadoop Distributed File System

HIV : Human Immunodeficiency Virus

ID3 : Iterative Dichotomiser 3

IDF : Inverse Document Frequency

IMBD : Internet Movie Database

NoSQL : Not Only Sql

RTF : Radyal Temelli Fonksiyon

S&P500 : Standard & Poor's 500 Index

Ses AB : Sosyoekonomik Statü Grubunda A Ve B'ye Dahil Olan Bireyler

SES : Sosyoekonomik Statü

SNNS : Stuttgart Neural Network Simulator

SQL : Structured Query Language

TD-IDF : Term Frequency–Inverse Document Frequency

TN : True Negative

TP : True Positive

TUİK : Türkiye İstatistik Kurumu

TV : Televizyon

VBA : Visual Basic for Applications

YSA : Yapay Sinir Ağları

GİRİŞ

Veri bilimi kavramını 2001 yılında "Data Science: An action Plan for Expanding the Technical Areas of Field of Statistic" adlı çalışması ile literatüre kazandırdığı düşünülen istatistikçi William Cleveland, makalesinde istatistiğin temel alanlarını genişletmek amacıyla bir plan önerdiğini ve bu değişen alanın adının "Veri Bilimi" olacağını belirtmiştir. 2000 yılında Ohsumi de "Veri Bilimi" ismini kullanmadan benzer konulardan bahsetmiştir. İki bilim adamı da veri analizi kavramının veri bilimine dönüşmesi gerektiğinin ve sadece istatistiksel metotların ve veri madenciliği yöntemlerinin veriden anlam çıkarmakta yeterli olmadığının üzerinde durmuşlardır. Dolayısıyla veri bilimini, istatistiksel modeller, metotlar, teoriler ve bilgisayar gücünü bir araya getiren çok disiplinli bir alan olarak öngörmüşlerdir. Harvard Business Review gibi kuruluşlar 21. Yüzyılın ve geleceğin en değerli mesleği olarak "Data Scientist-Veri Bilimcisi"ni seçmiştir. Bu da veri bilimi / analitiği disiplinlerinin günümüzdeki önemini ortaya çıkartan bir başka gözlemdir (Davenport ve Patil, 2012).

Klasik istatistik kavramında yapısal olan veri ve bu veri tipine uygun araçlar kullanılmaktadır. Fakat gelişen internet ve bilgisayar teknolojileri bilinen veri ve veri analizi kavramlarının değişmesine neden olmuştur. Teknolojideki gelişmeler verinin boyutunda ve çeşitliliğinde önemli bir artış yaratmıştır. 2013 yılında yapılan bir araştırmada, 2013 yılına kadar dünyada biriken verinin %90'ının son 2 yılda üretildiği belirtilmiştir (Dragland, 2013). Mobilite ve internet kullanımının yaygınlaşması veri boyutundaki ve çeşitliliğindeki artışta önemli rol oynamaktadır. Öte yandan başka bir çalışmada 2015 yılında küresel veri hacminin 7.9 zettabayt olduğu ifade edilmiştir. 1 zettabayt'ın 1 trilyon gigabayt olduğu düşünülürse veri boyutunda gelinen nokta daha net anlaşılabilmektedir. Erişilmesi, toplanması imkânsız gibi görünen bu veri boyutu aşağıdaki örnekler ile daha anlaşılır hale gelmektedir;

- New York Borsasında hisse senedi alışverişleri nedeniyle günlük 1 terabaytlık (1024 gigabayt) veri birikmektedir.
- Bir jet uçağı her 30 dakikada 10 terabaytlık veri toplamaktadır.

- Google günlük olarak 24 petabaytlık (1 petabayt = 1024 terabayt) veri işlemektedir.
- Facebook'da her 1 dakikada yaklaşık 34722 beğeni yapılmaktadır.
- Twitter'da günlük yaklaşık 175 Milyon ileti paylaşılmaktadır.
- Youtube'a her gün 65 bin yeni video yüklenmektedir.
- Whatsapp'da günde 27 milyar mesaj işlemektedir.
- Almanya Arjantin maçı bittiği an 1 dakika içerisinde 618,715 tweet atılmıştır.
- 4 Ekim 2012 yılında Barack Obama ve Mitt Romney arasında yapılan münazara sırasında 2 saat içinde 10 milyondan fazla tweet atılmıştır.
- Flickr adlı fotoğraf paylaşım sitesinde 2012 yılında Şubat'tan Mart'a kadar paylaşılan günlük ortalama fotoğraf sayısı 1.8 milyondur. Bu da her fotoğrafın 2 MB olduğu varsayıldığında günlük 3.6 terabayt'lık veri paylaşıldığı anlamına gelmektedir (Michel, 2012).

Öte yandan Uluslararası Veri Kuruluşu'nun yaptığı Dijital Evren araştırmasında şuan bile oldukça büyük boyutlarda olan veri miktarının, 2020 yılına kadar yıllık %44 oranında artacağı öngörülmektedir (Gantz ve Reinsel, 2012). Yukarıdaki örneklerde de görüldüğü üzere biriken verinin çeşitliliği ve düzensizliği de artmaktadır. Başlıca veri kaynakları olarak, makine verisi, üçüncü parti yazılımlar, mobil uygulamalar, akıllı okuyucular, sensörler, satış ve faturalama, sosyal medya, web sitesi kullanımı, müşteri ilişkileri yönetim sistemleri ve kurumsal kaynak planlama sistemlerinden bahsedilebilir. Bu kaynaklardan yapısal, yapısal olmayan ve yarı yapısal adı verilen 3 tip veri elde edilebilmektedir. Ürün stok tablosu gibi satır ve kolon bazındaki veriler yapısal veri tipindedir. Twitter ya da internet günlüklerinden elde edilen veriler gibi okuması ve anlamlandırması yapısal veriler kadar kolay olmayan veri tipine de yapısal olmayan veri denmektedir.

Verinin niceliğinin yanı sıra niteliğinin de katlanarak artması "Büyük Veri" kavramını ortaya çıkartmıştır. Büyük veri genel anlamıyla; yapısal, yarı yapısal ya da yapısal olmayan, büyük depolama alanları gerektiren, bilinen yazılım araçları ile makul

zamanlarda işlenmesi zor olan verileri ifade etmektedir. Büyük veri kavramı 1998 yılında ilk olarak Silicon Graphics şirketinin yayınlarında "Big Data and the Next Wave of InfraStress" başlığı ile karşımıza çıkmaktadır. Kavramın kullanıldığı ilk akademik çalışma ise 2000 yılının Ağustos ayında gerçekleşen Dünya Ekonometri Kongresi'nde Diebold Tarafından sunulan "Big Data Dynamic Factor Models For Macroeconomic Measurement And Forecasting" adlı bildiridir.

Gelişen teknoloji ile veri akışındaki artışın veri çöplüğü oluşturduğunu düşünmektense, bu verileri analiz ederek anlamlı bilgiler elde edilebilen değerli bir veri yığını olduğunu düşünmek büyük verinin bugün algılanan konumuna getirmiştir. Geçmişte birçok kişinin düştüğü veri çöplüğü yanılgısı 2010'lu yıllardan sonra ortadan kalkmaya başlamıştır. Birçok kurum veri saklamanın ve bu verinin analiz edilip anlamlandırılmasıyla kazanacağı rekabet avantajının farkına varmış ve buna göre strateji geliştirmeye başlamışlardır. Günümüzde internet devrimi ile işletmeler anlamlı bilgiler üretebilmek için sadece satış ya da üretim verilerini değil, aynı zamanda sosyal medya ve mobil cihazlardan elde edilen verileri de kullanmaktadır. Özellikle pazarlama ve satış alanlarında tüketici görüşlerini bilmek ve bunun üzerine strateji geliştirip karar almak önemli olduğu için işletmelerin ürün ya da hizmetleri hakkında sosyal medyadan ve mobil cihazlardan elde ettikleri verilerin analizi ve anlamlandırılması önem kazanmıştır. Belirli bir işletme, marka, ürün ya da hizmet hakkında önemli bir büyük veri kaynağı olan sosyal medyadan elde edilen verilerin nicelik ve nitelik olarak analiz edilmesi sonucunda değer yaratılması mümkündür. İletilerin niteliğinin ortaya çıkarılması amacıyla yapılan analizlerden ekseriyetle kullanılanlarından birisi de duygu durum analizidir. Böylece hakkında araştırma yapılan konuda kişilerin olumlu, olumsuz ya da nötr düşüncelere sahip olup olmadığı kolaylıkla ortaya çıkarılabilmektedir. Bu da sosyal medya madenciliği kavramının gelişmesinde ve önem kazanmasında önemli bir rol oynamaktadır.

Bu tez çalışmasında büyük veri analizi ile işletmelere fayda sağlanacak sonuçların çıkarılması amaçlanmıştır. Bu doğrultuda, veri kaynağı olarak sosyal medyadan yararlanılıp, işletme olarak bir televizyon kanalı seçilmiş ve bu işletmenin tüketicisi

konumunda olan seyircilerin görüşlerinin elde edilebilmesi için Twitter'dan ilgili kanal ve yayın akışındaki programlar hakkında veri toplanmıştır. Çünkü binlerce kişiye anket yapılarak elde edilmesi çok güç olan anlık seyirci görüşleri sosyal medya sayesinde kısa bir sürede elde edilebilmekte ve bunlardan değer yaratılabilmektedir. Duygu analizi ile seyircilerin kanal ve program hakkında yazdıkları iletilerin olumlu, olumsuz ya da nötr olarak sınıflandırılan duygulardan hangisini içerdiği incelenmiştir. Dakikada atılan tweet sayısı ve ilgili dakikada atılan tweetlerin duyguları ile oluşturulan veri setiyle kümeleme analizi yapılmıştır. Bu analiz sayesinde tweetlerin nitelik ve nicelik olarak incelenerek gruplanması sağlanmış ve bu gruplarla mevcut reyting sistemi ile ölçülen dakikalık reytinglerinin kategorize değerleri arasında ilişki olup olmadığı incelenmiştir. Son olarak da yayın akışındaki programların kendilerine has özellikleri ile yayın süresince aldıkları tweetler ve bunların duygularının girdi olarak kullanıldığı gelecek programın reyting değerinin tahmininin yapılmasını amaçlayan model kurulmuştur. İlgili model karar ağacı, destek vektör makineleri ve yapay sinir ağları yöntemleri kullanılarak çözülmüş ve modeller tahmin performanslarına göre karşılaştırılmıştır. Böylece kanala test yayını sürecinde olan program hakkında alacağı kararlara destek mekanizması kurulması hedeflenmiştir.

BİRİNCİ BÖLÜM

BÜYÜK VERİ

1.1. Büyük Veri Nedir?

Büyük veri kavramı için yapılabilecek tanımlamalardan bir tanesi; verilerin mevcut ilişkisel veri tabanları ile kolaylıkla yönetilemeyecek boyutlarda ve sürekli büyümeye devam eden bir yapıda olmasıdır. Ayrıca büyük veri, büyük boyutlardaki veri kümelerinin analizi ile gelecek için anlamlı bilgi çıkarılmasına olanak sağlamaktadır. Bu sayede değersiz görünen milyarlarca veriyi belirli bir yapıda ilişkilendirerek, anlamlandırmak ve kaliteli bilgiye dönüştürmek mümkün olmaktadır. Wu ve arkadaşları (2014) büyük veriyi bir grup kör adamın fili tarifine benzeterek açıklamışlardır. Bir grup kör adamdan önlerinde duran dev file dokunarak, dokundukları şeyin ne olduğunu tahmin etmeleri istendiğinde herkes kendi ulaşabildiği kadar yer hakkında bilgi sahibi olabileceğinden kişileri yanlış tahminlere sürükleyebilir. Büyük veri de aslında bu fil gibidir, büyük veri içerisinden küçük örnekler alınarak yapılan analizler yanlış sonuçlara götürebilir, bu nedenle verinin bütünüyle ele alınması çok önemlidir.

Uluslararası Veri Kuruluşu ise büyük veriyi; verinin kendisi, bu verinin analitiği ve bu analitiğin sonuçlarının işletmelerde değer yaratabilecek şekilde sunulması olarak tanımlamaktadır. Büyük veri disiplinler arası bir kavram olduğu için derlenmesi, saklanması ve çözüm gücü için teknolojiye, bu verilerin analizi ile çıkarılacak sonuçların değer bulabilmesi, yorumlanabilmesi ve uygulanabilmesi için de sosyal ve ekonomik alanlara birlikte ihtiyaç duymaktadır. Özellikle teknolojideki gelişmeler ile veri kaynakları değişmiş ve dolayısıyla verinin boyutu da değişmiştir; uydular, sensörler, sosyal medya, GPS sinyalleri ve mobil cihazlar önemli büyük veri kaynaklarındandır. Büyük veri analizi sadece satır bazlı yapısal verilerle değil çok çeşitli kaynaklardan elde edilebilecek yarı yapısal veya yapısal olmayan veriler ile de analiz yapılmasını sağlamaktadır. Öte yandan örneklem verisinin analizinin yeterli olmayacağı durumlarda, büyük veri ile açıklayıcı ve keşfedici analizlerin yapılması mümkün olmaktadır.

Dolayısıyla büyük veriyi; yeni bilişim teknolojilerinin araç ve yapılarının yardımı ile çeşitli veri kaynaklarından verilerin toplanması, saklanması, düzenlenmesi ve analiz edilip sosyal ve ekonomik anlamda değerli bilgiler ortaya çıkarılması olarak da ifade etmek mümkündür. Literatürde geçen büyük veri tanımlarından bazıları Tablo 1'de derlenmiştir.

Tablo 1: Büyük veri tanımları

Yazar, Tarih	Tanım
IBM, 2012b	Büyük Veri: Sensörler, sosyal medya paylaşımları, dijital resim ve görüntüler, işlem kayıtları ve cep telefonu GPS sinyalleri vb. kaynaklardan veri elde edilmesidir.
Johnson, 2012	Büyük Veri: Sensör çıktıları, çok büyük miktarlardaki müşteri davranışları, sosyal medya ve konum paylaşımlarıyla ilişkili veri setleridir.
Davenport v.d., 2012	Büyük Veri: İnternetten elde edilen tıklama dizilerinden, gen haritalarına kadar elde edilen her çeşit veridir.
Manyika v.d., 2011	Büyük Veri: Tipik veri tabanı yazılımlarının elde etme, toplama ve organize etme kapasitelerinin ötesinde büyüklüğe sahip olan veri setleridir.
Fischer v.d., 2012	Büyük Veri: Basit yöntemlerle işlenemeyen veridir.
Jacobs, 2009	Büyük Veri: İlişkisel bir veri tabanına yerleştirmek için çok büyük olan ve gelişmiş istatistik/görselleştirme programlarıyla analiz edilen veri paketleridir. Analiz edilmeleri için onlarca hatta yüzlerce paralel çalışan yazılıma ihtiyaç
Boyd and Crawford, 2012	Büyük Veri: kültürel, teknolojik ve akademik fenomeni karşılıklı etkileşimlerine dayanmaktadır. 1. Teknoloji: Büyük veri setlerinin bir araya getirme, ilişkilendirme ve karşılaştırma bilgisayar gücünün maksimize edilmesi ve algoritmik hassasiyet gösterilmesi. 2. Analiz: ekonomik, sosyal, teknik ve hukuki iddialar yapabilmek amacıyla büyük veri setleri üzerinde şablonların tespit edilmesi. 3. Mitoloji: Büyük veri setlerinden daha öncekinden daha kapsamlı çıkarımların yapılabileceği inancı.

Öte yandan literatürde büyük verinin tanımında 3V, 4V ve 5V şeklinde üç farklı yaklaşım yer almaktadır. Bu tanımların hepsi temelde 3V'nin geliştirilmiş hali şeklinde olup şöyledir (Şeker, 2015; Wamba v.d., 2015);

- 3V: Veri Büyüklüğü (Volume), Hız (Velocity), Çeşitlilik (Variety) (Gartner, 2012; Kwon ve Sim, 2013).

- 4V: Veri Büyüklüğü (Volume), Hız (Velocity), Çeşitlilik (Variety), Değer (Value) (Oracle,2012; Forrester, 2012).

- 5V:Veri Büyüklüğü (Volume), Hız (Velocity), Çeşitlilik (Variety), Değer (Value), Güvenilirlik (Veracity) (White, 2012).

Tüm tanımlarda ortak yer alan verinin **büyüklüğü** kavrama adını vermekte ve saklanılması için ihtiyaç duyulan alan ve verinin boyutunun büyüklüğünü ifade etmektedir. Objelerin birbirileriyle ve çevreleriyle iletişime geçebilmesi olarak tanımlanan Şeylerin İnterneti kavramının günlük hayatımızda yer alması üretilen verinin boyutunu hızla arttırmaktadır. Bunun yanı sıra araba, oyuncak ve beyaz eşya gibi birçok ürünün içine dâhil olan bilgisayar teknolojisi de ilgili ürünlerde bir çok veri üretilmesine ve bunun saklanabilir olmasına neden olmuştur (Erevelles ve ark., 2016). Öte yandan mobilizasyonun artması, kişisel bilgisayar, tablet ve telefonların yaygınlaşması, mobil ağların her an her yerde ulaşılabilir olması da üretilen verinin boyutunda büyük rol oynamaktadır. Milyonlarca kişinin kişisel cihaz sahibi olduğu düşünülürse, cihazlardan elde edilen GPS verileri, kullanılan uygulamaların biriktirdiği veriler ve kişilerin sosyal medyayı kullanarak ürettikleri veriler bile biriken verinin boyutunun büyüklüğünü açıklamaktadır. Endüstrideki gelişmeler, daha çok makinanın kullanılması ve bunların sürekli veri biriktiriyor olması da büyük verinin oluşmasına neden olmaktadır. Sadece üretim değil hizmet sektörü ile toptan ve perakende satış alanında da müşteri bilgileri, satış rakamları ve stok- ürün bilgileri gibi verilerle büyük veri oluşmaktadır.

Hız özelliği, büyük verinin hızla oluşması ve verinin analizinde hızlı olmaya ihtiyaç duyulmasını ifade etmektedir. Akıllı cihazlar ve sensörler gibi dijital cihazların yaygınlaşması geçmişte benzeri görülmemiş bir süratle veri üretilmesine neden olmaktadır. Bu hızlı veri akışı karşısında anlık-gerçek zamanlı analizlere olan ihtiyaç artmaktadır. Gerçek zamanlı analizler sayesinde hızla akan verinin aynı hızla analiz edilip anlamlı bilgiler elde edilmesi ve stratejiler geliştirilmesi mümkün olmaktadır. Örneğin kişisel cihazlardan kişilerin alınan bilgilerin gerçek zamanlı analizleri ile tüketicilerin ilgileri doğrultusunda anlık ve kişiye özel fırsatlar sunmak mümkündür. Bunu günümüzde

birçok arama motoru, video paylaşım sitesi veya alışveriş sitelerinde görmek mümkündür. İlgili sayfalarda yaptığınız arama, incelediğiniz ürün veya izlediğiniz videolara bağlı olarak size beğenebileceğiniz ürün ya da videolar önerebilmektedir. Geçmişteki benzer durumları inceleyip yeni kişinin girdilerini gerçek zamanlı analiz eden bir algoritma geliştirerek müşteri üzerinde değer yaratılması hedeflenmektedir. Saniyede binlerce metin paylaşılan Twitter ya da binlerce dakikalık video paylaşılan Youtube gibi sosyal medya kanalları da büyük verinin hızla üretildiği alanlardır. Buralardan anlık verinin toplanabilmesi mümkün olduğu için verinin akış hızına uygun sistemlerin geliştirilmesi ile anlık analizlerin yapılması mümkündür.

Çeşitlilik, büyük veriyi geleneksel veri kavramından ayıran özelliklerden biri de büyük verinin, geleneksel verilerdeki gibi sadece yapısal verilerden oluşmamasıdır. Büyük veri yapısal, yapısal olmayan ya da yarı yapısal biçimde olabilmektedir. Veri kaynağındaki çeşitlilik üretilen verinin de format ve içerik bakımından yüksek çeşitliliğe sahip olmasına neden olmaktadır. İlişkisel veri tabanı, web sayfaları, makine verileri, sensörler gibi veri kaynaklarının her biri farklı yapıda veri üretmektedir. Yapısal veri: günümüzde biriken verinin yaklaşık %5'ini oluşturan (Cukier, 2010), ilişkisel veri tabanlarında ve çalışma sayfalarında derlenen tablosal verilerdir. Yapısal veriler, üzerinde terim olarak işlem yapılması mümkün verilerdir. Metin, fotoğraf, ses kaydı ve video gibi çeşitli veri kaynaklarından derlenebilen ve ilişkisel veri tabanları ile yönetilemeyen verilere yapısal olmayan veriler adı verilmektedir. Bu gibi yapısal olmayan verilerin analiz edilebilmesi için yapısal bir düzenleme ile dönüştürüldüğü haline ise yarı- yapısal veriler denmektedir. Algoritmalarda kullanılabilmesi amacıyla kullanıcı tarafından etiketlenmiş, genişletilebilir biçimlendirme dili (XML), sosyal medya yayınları, e-postalar, web sayfası günlükleri gibi veriler yarı- yapısal veri tipine örnektir (Jing ve ark., 2015)

Büyük veri analizinde önemli bir unsur olan veriden ilgili kurum ya da kişilere yarar sağlayacak bilgilerin elde edilmesi, **değer** özelliği ile açıklanmaktadır. Büyük veri işletmeler için yatırım geri dönüşü bakımından azımsanmayacak bir hıza sahiptir bu nedenle yarattığı değer oldukça fazladır. Verinin karar verme süreçlerine anlık olarak

incelenip, değiştirilip, değerlendirilmesi, doğru kararların alınmasında önemli rol oynamaktadır (Ünal, 2015). Bu da birçok organizasyon için katma değer anlamına gelmektedir. Müşteri ihtiyaçlarının önceden tahmini, bir aracın anlık takibi, video kayıtlarından yüz tanıma, kredi verilecek müşterinin demografik özelliklerinin yanı sıra en çok alışveriş yaptığı yeri, en çok gittiği restoranı da bilme, sosyal medyaya yazılan metinlerden kişilerin duygularını analiz etme, müşterileri belirli özelliklerine göre kümeleyip ona göre pazarlama stratejisi geliştirme, ses kayıtlarından yalan söyleyip söylenmediğini analiz etme gibi çeşitli alanlarda değer yaratılabilmektedir.

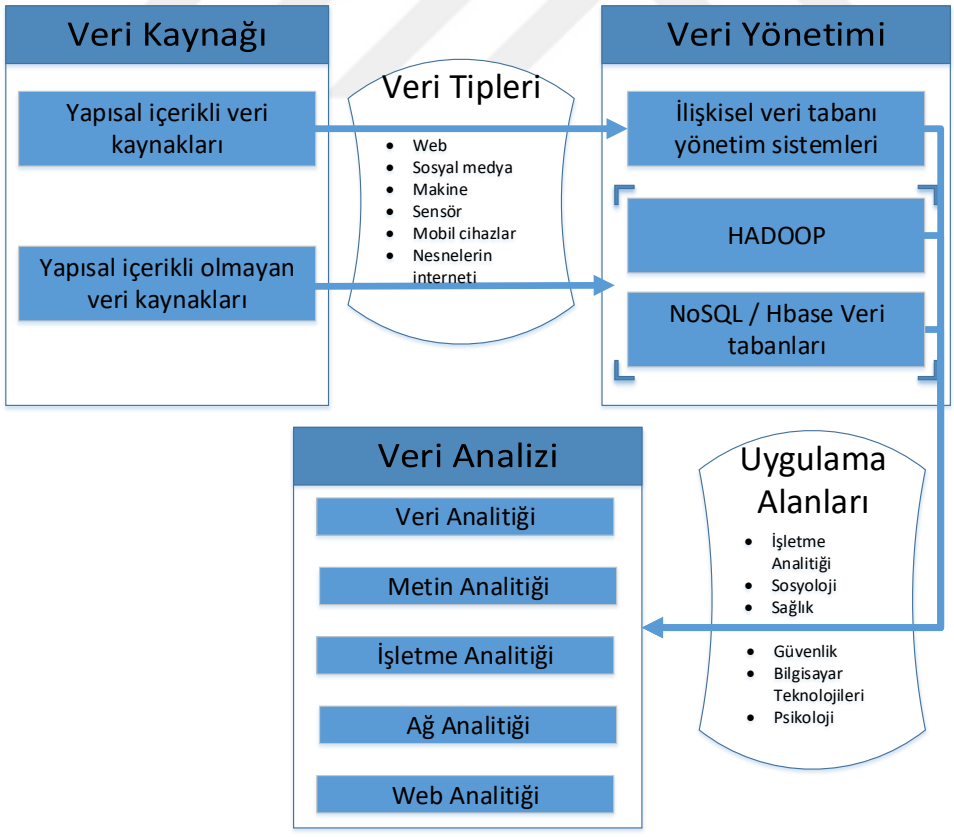
Son özellik olan **güvenilirlik** (kimi kaynaklarda doğrulama (vertification) olarak geçmektedir), her verinin ve her analiz sonucunun kayıtsız şartsız geçerli olmadığını, verinin elde edildiği kaynağın, akışının ve yapılacak analizin önem arz ettiğini belirtmektedir. Elde edilen veri güvenilir ve doğru olmalıdır. Bunun için yapılan analizlerin değerlendirilmesinde kullanılan doğruluk, kesinlik ve duyarlılık ölçütlerinin sonuçları önemlidir. Veride uyumsuzluk, anlam karmaşası, hile yapılması, tekrarlanma, eksiklik ve gizlilik gibi şeyler kesin, doğru ve güvenilir olmayan analizlere neden olmaktadır. Büyük verinin sonuçlarının kanıtlanması mümkün olmasa da güvenilirlik özelliği sayesinde sonuçlar belli bir olasılık ile geçerliliğe sahip olabilmektedir (Ermani ve ark, 2015).

Tablo 2: 5V tanım ve örnekleri

V	Tanım	Örnekler
Büyüklük	Çok büyük saklama alanı sarf eden veya çok fazla sayıda kayıt içeren, büyük hacimli veri setleri.(Russom, 2011)	Wal-Mart'ın veri deposu 2.5 petabyte'dan fazla bilgi içermektedir (Manyika v.d., 2011)Dell satış ve pazarlama faaliyetlerine ilişkin olarak 1.5 milyon kayıttan oluşan bir veri tabanı oluşturmayı hedeflemektedir (Davenport, 2006). Tesco'da ayda 1.5 Milyardan fazla veri üretilmektedir (Manyika v.d. 2012)
Çeşitlilik	Çeşitli kaynaklardan, farklı formatlarda ve çok boyutlu veri alanlarında üretilen veri setleri. (Russom, 2011)	Procter&Gamble, fonksiyonel alanlar arasındaki karşılıklı ilişkileri inceleyerek işletme performansını arttırmak amacıyla; operasyon, tedarik zinciri, satış, müşteri araştırmaları ve pazarlama alanlarında çalışan 100'den fazla analistten oluşan bir grup oluşturmuştur. Tata Motor, her ay müşteri şikâyetlerinden servis kayıtlarına kadar 4 milyondan fazla metni analiz ederek müşterilerinin memnuniyet seviyeleri arasında sürekli bir bağ kurmaktadır.
Hız	Veri üretiminin ve/ya dağıtımının frekansı. (Russom, 2011)	Amazon, yeni ürünleri, tedarikçileri, müşteri ve promosyonlarıyla; temin ettiği teslim tarihlerini aksatmadan sürekli bir akış sağlamaktadır. (Davenport, 2006)Perakendeciler her bir müşterilerinin verilerini internetten elde ettikleri tıklanma akışları da dâhil olan çok farklı olan kaynaklardan takip edebilmektedir. Ek olarak perakendeciler müşterilerinin davranışlarını neredeyse eş zamanlı olarak görebilmekte sürekli artan oranlarda güncelleyebilmektedir.
Güvenilirlik	Doğası gereği tahmin edilmesi güç olan büyük verinin analiz edilerek daha güvenilir tahminler elde edilmesi.	eBay Inc. farklı kaynaklarda oluşan aynı verinin 20 ila 50 sefer yinelenmesinden kaynaklanan çok büyük bir veri tekrarı sorunu yaşamaktadır. Bu sorunu çözmek için eBay, yöneticilerin veri tekrarını filtreleyebileceği dahili bir internet sitesi geliştirmiştir.
Değer	Verinin işlenmesi ve dönüştürülmesiyle ekonomik açıdan değerli, faydalı olması.	Premier Healthcare Alliance şirketi veri analitiği kullanarak, harcamalarında 2.85 milyar dolarlık azalış ve müşteri sayısında artış elde etmiştir (IBM, 2012a)

Tablo 2'de özetlendiği ve örneklendiği üzere büyük veri; nicelik ve kapladığı alan anlamında büyük, hızlı oluşan ve işlenmesi için hız gerektiren, çeşitlilik içeren, güvenilir ya da doğrulanabilir olan ve verinin içerisindeki bilginin ortaya çıkarılması ile değer yaratan bir kavramdır. Öte yandan büyük veri sadece matematiksel analizler olarak düşünülmemelidir. Büyük veri, daha çok yeni nesil bilgi teknolojileri imkânlarından faydalanılarak, verinin farklı kaynaklardan toplanması, organize edilmesi ve analitik olarak incelenerek değer yaratan iç görü üretilmesi ve bunların rekabetçi avantaj yaratmak üzere paydaşlarla paylaşılmasıdır (Şekil 1). Böylelikle büyük veriyi, 5V'nin yönetilmesi, işlenmesi ve analizi ile sürdürülebilir değer yaratan, performans ölçen ve rekabetçi avantaj elde edilmesini sağlayan bütünleyici bir yaklaşım olarak tanımlayabiliriz.

Şekil 1: Büyük veri mimarisi



Kaynak: Chen v.d., 2012; Hashem v.d., 2015

1.2. Büyük Veri Teknoloji Ve Araçları

Büyük verinin niteliği ve niceliği nedeniyle, saklanması, işlenmesi ve anlamlandırılması aşamalarında geleneksel veri tabanları ve işleme teknikleri yetersiz kalabilmektedir. Büyük veri, bilgiye dönüştürülmesi amacıyla geliştirilen birçok platform sayesinde, yeni bir teknolojik destek sektörü oluşturmuştur (Sağıroglu ve Sinanç, 2013). Büyük veri kavramının gelişmesi ve önem kazanması ile ilişkisel veri tabanı yönetim ve işleme sistemlerine alternatif ya da destekleyici yeni nesil sitemler geliştirilmiştir.

İlişkisel veri tabanları, verilerin satır ve sütunlar halinde düzenlenmiş tablolarda tutulduğu ve bu tablolar arasında ilişkilerin tanımlanabildiği veri depolama sistemleridir. Farklı tablolarda biriktirilen veriler çeşitli anahtar kelimeler aracılığı ile birbirine bağlanabilmektedirler. Böylece ilgili anahtar kelimelerle yapılan sorgularda çeşitli tablolardaki veriler bir arada görülebilmektedir. İlişkisel veri tabanı sistemlerinin en çok kullanılan kavramlarından birisi SQL'dir. SQL kavramı yapılandırılmış sorgu dili anlamına gelmekte ve veri tabanındaki verilerin yönetilmesinde kullanılmaktadır. MySQL; açık kaynak kodlu, Windows ve UNIX gibi platformlarda ücretsiz sunulan, bu nedenle milyonlarca sistemde yüklü olan en çok tercih edilen ilişkisel veri tabanı yönetim sistemlerinden birisidir. Veri boyutlarının artması karşısında ilgili yazılımın da geliştirilmesiyle, milyonlarca kayıt içerebilen binlerce tablodan oluşan veri tabanlarının saklanmasında halen MySQL kullanılabilmektedir. PostgreSQL de MySQL gibi açık kaynak kodlu ve ücretsiz kullanım imkânı sunan bir başka veri tabanı yönetim sistemidir. Bunların dışında benzer işlevlere sahip, birçok ticari lisanslı veri tabanı sistemleri de bulunmaktadır.

Büyük veri setlerinde depolama ve veri işlenmede geliştirilen yeni nesil platformların en önemli temsilcilerinden birisi Hadoop'tur. Hadoop dağıtılmış dosya sistemleri, analitik ve veri depolama platformları içeren paralel hesaplama ortamıdır (Zikopoulos ve Eaton, 2011). Açık kaynak kodlu bu yazılım projesinde Yahoo, Microsoft, Cloudera, Twitter, IBM gibi teknoloji firmaların katkıları bulunmaktadır. Hadoop, kullandığı basit programlama dili ile büyük hacimli verileri kümelenmiş bilgisayarlar arasında dağıtarak

işlenmesini amaçlayan bir yazılımdır. Binlerce sunucuyu barındıran kümeler üzerinden çalışması sayesinde hızla büyüme olanağına sahip olması büyük hacimli veri setleri ile çalışanların ilgisini çekmektedir. Genel hatlarıyla Hadoop, ölçeklenebilir (kullanıcının ihtiyaç duyduğunda yeni düğüm ekleyebildiği), hesaplı (yüksek donanım maliyetleri gerektirmeyen), veri tipi ve kaynak bakımından esnek (yapısal ya da yapısal olmayan her türlü veri setinin ve veri kaynağının kullanılabildiği) ve hata esnekliği olan (herhangi bir düğümde sorun çıkması durumunda kümedeki diğer sunucular ile çalışmanın sürdürülebildiği) bir yazılımdır.

Hadoop kümeleri arasındaki iletişimi sağlamak ve çeşitli girdi ve çıktı düğümlerindeki dosya sistemlerini birbirine bağlayarak büyük bir dosya sistemi yaratmak amacıyla HDFS yani dağıtık Hadoop dosya sistemi geliştirilmiştir (Ünal, 2015; Schneider, 2012). HDFS, birçok sunucu diskinin bir araya getirilmesi ile sanal bir disk yaratılmasını sağlar. Böylece büyük boyutlardaki dosyalar bu sanal diskte saklanabilmekte ve okunabilmektedir. HDFS'de saklanan büyük hacimli verilerin işlenebilmesi amacıyla EşleİndirgeEşleİndirge (MapReduce) yöntemi kullanılmaktadır. EşleİndirgeEşleİndirge, kümeler üzerine dağıtılan işlemlerin eş zamanlı işlendiği ve sonrasında tekrar birleştirildiği güçlü bir paralel programlama tekniğidir. Bunların dışında Hadoop platformunda kullanılan diğer yöntem ve sistemlerden bazıları; HBase, Pig, Hive, Sqoop ve Oozie'dir (Zikopoulos ve Eaton, 2011). Hbase, rassal yazma ve okuma erişimi olan ölçeklenebilir dağıtık bir veri tabanı sağlamaktadır. Pig, veri setlerinin analizini sağlayan yüksek kaliteli bir veri işleme sistemidir. Hive, SQL şeklinde ilişkisel model, veri depolama ve sorgulama sağlayan bir veri depolama çözümüdür. Sqoop ise ilişkisel veri tabanlarından Hadoop'a veri aktarılmasını sağlayan bir projedir. Oozie, bağımsız Hadoop işlemleri için iş akışı sağlayan bir sistemdir.

Hadoop gibi ilişkisel veri tabanı sistemine alternatif olarak geliştirilen ve uygulamada sıklıkla kullanılan pek çok sistem geliştirilmektedir. Bunlardan NoSQL de Hadoop gibi dağıtık veri depolama yöntemi ile ilişkisel veri tabanlarının büyük hacimli veri setlerindeki yetersizliklerine çözüm olmayı amaçlamaktadır. MongoDB'de ölçeklenebilir,

açık kaynak kodlu ve genellikle sürekli büyümekte olmayan veri setleri için geliştirilen NoSQL veri tabanı çalışmasıdır. İlişkisel veri tabanlarından farklı olarak büyük hacimli verilerde hızlı okuma, yazma ve sorgulama özelliklerine sahiptir.

Oluşan ve biriken verinin her geçen gün artması ve bunların bilgi çıkarılabilecek değerli unsurlar olduğunun anlaşılması, yeni nesil veri depolama ve işleme teknolojilerinin geliştirilmesi ihtiyacını doğurmuştur. Dolayısı ile gelişen internet ve bilgisayar teknolojisinin veri hacminde büyümeye neden olduğu gibi ortaya çıkan depolama ve işleme ihtiyacına da çözüm sunduğu söylenebilir.

1.3. Büyük Veri ve İşletme Analitiği

Günümüzde iç içe geçmiş kavramlar olan veri analitiği ve istatistikte ortak amaç toplanan verilerin işlevsiz unsurlar olarak görülmemesi ve bunların analiz edilerek anlamlı sonuçların ortaya çıkarılması ile işletmeler için değer yaratmaktır. İşletme analitiğinin temel konularından biri veri seti içerisindeki ilişkilerin açığa çıkartılmasıdır. Bu amaçla yapılan veri analizi çalışmalarının temelinde istatistik vardır (Provost ve Fawcett, 2013). İstatistik, geliştirdiği teorilerin/hipotezlerin gerçek verilere ve problemlere uygulanması ile yararlılığı ortaya konabilen, disiplinler arası, yöntemler bütünüdür. İstatistiksel analizlerde kullanılan iki temel veri toplama yöntemi tamsayım ve örneklemedir. Ana kütlenin tamamından veri toplanırsa tamsayım, rassal olarak seçilmiş bir gruptan veri toplanırsa örnekleme yöntemi yapılmış olmaktadır. Ana kütlenin tamamına ulaşılmasının güç ya da maliyetli olduğu durumlarda ana kütle hakkında çıkarımların örnekleme verilerine dayandırılması istatistiksel çıkarımdır. İstatistiksel yaklaşım öncelikle araştırma sorusunun belirlenmesini, buna göre hipotezlerin geliştirilmesini, ardından verinin toplanması ve kuramın geçerliliğinin araştırılması üzerine kurulmuştur (Dhar, 2013). İstatistikte hipotezden kurama ulaşmayı hedeflerken, veri ve hesaplama yöntemi önem bakımından karşılaştırıldığında hesaplamaya daha çok önem verildiği görülmektedir. Gelişsen teknoloji bilimsel yaklaşımları, veri setlerine göre araştırma sorularının geliştirildiği, ilgili veri setlerinden ne tür ilişki ve korelasyonların var olduğunun araştırıldığı, hipoteze dayalı yöntemlerden keşfetmeye dayalı yöntemlere doğru

kaydırmaktadır (Gürsakal, 2014). Geçmişte araştırma sorularına göre düzenlene deneyler ile veri toplanırken, günümüzde gelişen teknoloji ile dakikada binlerce verinin toplanması mümkün olmaktadır. Veri boyutlarının her geçen gün büyümesi, büyük veri kavramının ortaya çıkması geleneksel yöntemlerin uygulanabilirliğinin sorgulanmasına neden olmaktadır (Pentland, 2012). Büyük veri kavramının içerdiği veri seti, bilinen örnekleme ve tamsayım yöntemlerine uymamaktadır. Büyük veri, iki kavram arasında tamsayıma daha yakın bir noktada yer almaktadır ve haliyle bu yeni kümeden ana kütle değerlerine dair çıkarım yapılabilecek bir kuramsal bilgi bulunmamaktadır. Bunun yanı sıra verinin boyutu büyüdükçe hipotez testlerindeki her soru istatistiksel olarak anlamlı çıkmaya meyilli olmaktadır (Anderson, 2008).

Bilgisayar teknolojisindeki gelişim büyük verinin ortaya çıkmasına sebep olurken aynı zamanda yeni analiz tekniklerinin geliştirilmesini de sağlamıştır. Teknolojik yenilikler sayesinde yeni hesaplama ve analiz tekniklerinin de ortaya çıkması verinin başlangıç noktası olarak alındığı yöntemlerin uygulanmasını daha mümkün kılmıştır. Veri analitiği teknikleri, verinin önemli olduğunu savunan akımla gelişmiştir. Temelinde istatistiksel yöntemleri de barındıran bu teknikler, veriden bilgi keşfetmeyi amaçlamaktadır. Verinin hacim ve çeşitlilik bakımından büyümesi, kullanılmakta olan birçok tekniğin yetersiz kalmasına neden olmuştur. Fakat aynı dönemde daha güçlü bilgisayarların ortaya çıkması, veri setinin bütününü kullanabilen, geçmişte yapılanlardan daha derin analizlerin yapılabildiği algoritmaların geliştirilmesi bu verilerin işlevsiz yığınlar olmadığını ortaya koymuştur. Böylece veri analitiği kavramının gerek çözüm teknikleri gerek veri toplama ve depolama teknikleri bakımından geliştirilmesi ile büyük veri analitiği kavramı ortaya çıkmıştır. Bu yöntemlerin birbirini net olarak kapsadıkları söylenemese de birbirilerinden net bir şekilde ayrılması da mümkün olamamaktadır (Şekil 2).

Şekil 2: İstatistik, veri analitiği, büyük veri analitiği

IDC, büyük veriyi tanımamlarken; verinin kendisinin, verinin analitik olarak incelenmesinin ve bu inceleme sonuçlarıyla işletmelerde değer yaratmasının önemli olduğunu belirtmiştir (Gantz ve Reinsel, 2012). İşletme analitiğinin temelini de bu üç unsur oluşturmaktadır. İşletme analitiği genel tanımı itibariyle; işletmelerin içyapısında biriken veya işletmeler için önemli olan dış kaynaklardan toplanan verilerin derlenmesi, depolanması, düzenlenmesi ve analiz edilmesiyle anlamlı bilgiler ortaya çıkararak, işletmeler için faydalı yaratılmasıdır. Diğer bir deyişle işletme analitiği, veri analitiği aracılığıyla vakaların anlaşılması için kullanılan prensipler, süreçler ve teknikler bütünüdür. Veri temelli kararlar, tecrübeler ve sezgiler yerine verinin analiz edilmesi ile alınan kararlardır. Bu sayede işletmeler müşteri bağlılığından, eleman seçimine, satış tahmininden, yatırım kararına çeşitli alanlarda verilerin analiz edilmesiyle başarılı kararlar alabilmektedirler. Akademik ve sektörel birçok çalışma, işletmelerin büyük veri analitiği ile %15-20 arasında yatırım geri dönüş hızı kazandığını göstermektedir (Perrey v.d., 2013). Öte yandan McKinsey'ye göre büyük verinin toplanması, depolanması ve bilgi edinmek için analiz edilmesi; dünya ekonomisinde anlamlı bir etki yaratabilecek, üretkenliği ve rekabeti geliştirebilecek ve kamu alanında tüketiciler için ekonomik bolluk yaratabilecek bir süreçtir. (Manyika v.d., 2011).

İşletmelerin karlılıklarını arttırabilmeleri, rekabet avantajlarını kaybetmemeleri ve müşteri memnuniyetini sağlamaları gibi konular işletme analitiğinin çalışma konularıdır. Bu amaçla veri kaynakları, veri ambarları, veri tabanı yönetim sistemler kullanılarak, istatistiksel analiz, veri madenciliği ve veri görselleştirme gibi çeşitli analizler kullanılmaktadır. İşletme analitiği çalışmaları genellikle üç açıdan incelenmektedir; tanımlayıcı analitik, tahminleyici analitik, kuralcı analitik (Evans ve Lindner, 2012). Tanımlayıcı analitik, işletmelerin bütçe, satış, kazanç ve maliyet gibi verilerinin anlamlı grafik ve raporlarla özetlenmesidir. Dolayısı ile tanımlayıcı analitik geçmişte yaşanan olaylar hakkında bilgi vermektedir. Tahminleyici analitik ise işletmelerin satış rakamları, reklam kampanyalarının etkileri gibi veri setlerinden geçmiş olaylar hakkında bilgi edinip gelecek için tahmin üretmeyi amaçlayan çalışmalardır. Diğer bir deyişle geçmiş verideki örüntüleri ve ilişkileri ortaya çıkararak bunların gelecek zamanlar hakkında fikir yürütülebilmesi için anlamlandırılmasıdır. Kuralcı analitik ise, işletmelerin üretim, satış, pazarlama ve finans alanlarında alacağı kararlarda, optimizasyon tekniklerini kullanarak, amaçları ve kaynakları doğrultusunda en iyi (optimal) ya da en iyiye yakın uygun (feasible) sonuçlar elde edebileceği modeller geliştirilmesini amaçlar. Kuralcı analitiği ve tahminleyici analitiğin sentezlenerek veri setindeki belirsizliğin de hesaba katıldığı çalışmalar da yapılabilmektedir.

1.4. Büyük Veri İle Veri Madenciliği

1.4.1. Sınıflandırma

Veri madenciliğindeki iki temel amaç, veri setindeki örüntülerden anlamlı bilgiler çıkararak gelecek için tahmin yapılması ya da veri setinin özelliklerinin tanımlanmasıdır. Bu amaçla geliştirilen modeller genel tanımları itibariyle “tahmin edici modeller” ya da “tanımlayıcı modeller” olarak adlandırılmaktadır (Akpınar, 2014). Veri madenciliği problemlerinde ulaşılması hedeflenen amaç önceden belirlenmelidir. Buna göre kullanılacak olan model seçilmeli veri seti ilgili modele hazır hale getirilmelidir. Sınıflama ve regresyon modelleri geçmiş veri setinin incelenmesi ve örüntülerinin değerlendirilmesi ile geleceğe yönelik politikalar geliştirebilme özelliğine sahip olmasından dolayı tahmin

edici modeller sınıfına dâhil olmaktadır. Kümeleme, birliktelik analizi, ardışık zamanlı örüntüler ise veri setinin genel yapısını gösteren hakkında bilgi veren tanımlayıcı modelleri oluşturmaktadır.

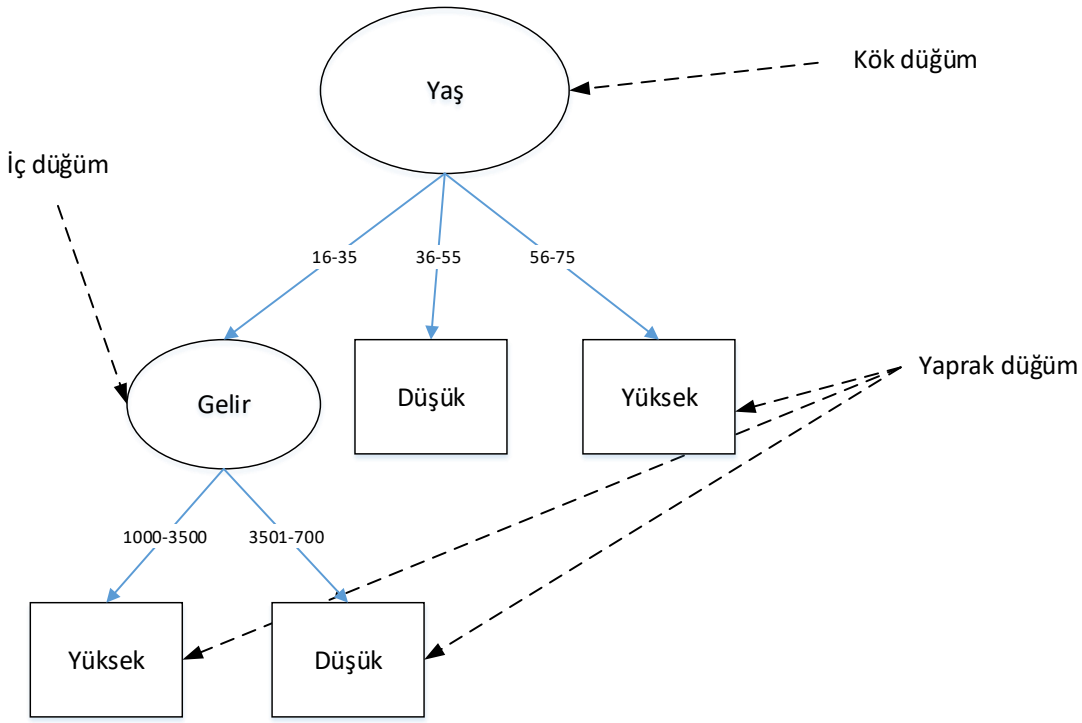
Tahmin edici modellerin çalışma prensibi insanların öğrenme prensibine benzetilebilmektedir. İnsan beynini veri tabanı olarak düşünürsek doğdukları günden itibaren 5 duyu organı ile algılayabildikleri her nesneyi öğrenerek veri tabanında depolarlar. Yeni karşılaşılan nesneler için ise veri tabanındaki var olan bilgileri değerlendirerek çıkarım yapmaya çalışır. Örneğin daha önce hiç çilek görmemiş bir insana, bu nesnenin ne olduğu sorulduğunda; koklayarak, tadarak ve şeklini inceleyerek geçmişte gördüğü meyve veya sebzelere benzetip bir çıkarım yapması beklenir. Bu durumda o kişi sınıflandırarak yeni nesneyi tahmin etmiş olmaktadır. Bu tahmin yapılmadan önce kişinin çileği domates ya da bambaşka bir nesne olarak sınıflandırması durumunda kişiyi uyararak, çileğin ya da diğer nesnelerin özniteliklerin kişi tarafından algılanmasını sağlayarak yapılan öğrenme türüne denetimli öğrenme adı verilmektedir. Bu sayede kişi tahmin yürütmeden önce belirli özniteliklerin hangi değerleri aldığında hangi sınıfa ait olacağını bilmektedir ve buna göre gelecek için tahmin yürütebilmektedir. Fakat sınıf değeri hakkında hiçbir bilginin olmadığı, sadece öznitelik değerlerine göre yapılan sınıflandırma çalışmaları denetimsiz öğrenmedir.

Veri madenciliğinde, denetimli öğrenmede, veri seti sınıf etiketi var olan eğitim seti ve sınıf etiketi yer almayan test seti olarak ikiye ayrılır. Kurulan veri madenciliği modeli sayesinde, eğitim seti ile veri setinin bütünüyle ilgili örüntüler öğrenilir ve test setindeki verilerin sınıf değerlerinin tahmini ile öğrenme başarısı değerlendirilir. Öte yandan öğrenmiş olan bu model sayesinde yeni bir verinin hangi sınıfa ait olduğu da tahmin edilebilmektedir. Bu özellikleri ile sınıflandırma modelleri; pazarlama stratejisi kararları, finansal performans değerlendirme ve metin madenciliği gibi çeşitli konularda kullanılabilmektedir.

1.4.1.1. Karar Ağaçları

Karar ağaçları sınıflandırma algoritması, sürekli ve kategorik veri setleri ile çalışabilen denetimli öğrenme algoritmasıdır. Yaprakların sınıf değerleri, yapraklara ulaşılmasını sağlayan dalların ise öznitelikler ve onların aldığı değerlere benzetildiği gerçek bir ağaç yapısından esinlenerek geliştirilen bir algoritmadır. Bu nedenle karar ağacı algoritmalarında 3 ana düğüm bulunmaktadır; kök düğüm, ara/iç düğüm ve yaprak düğüm (Tan, 2006). Karar ağacında öğrenme süreçleri kök düğümünden yapraklara doğru yapılan bölme işlemleri ile gerçekleştirilir. Algoritmada yaprak düğüm hariç her düğümde bir kural tanımlanarak bu kurallar sonucunda ulaşılacak yaprak düğümün sınıf değerinin belirlenmesi hedeflenir. Örnek olarak banka müşterilerinin risklilik düzeylerinin, gelir, yaş ve mevcut müşteri olup olmadığı bilgilerine göre sınıflandırıldığı bir karar ağacı Şekil 3'de gösterilmektedir. Örnekte kök düğüm olan yaş özniteliğinden başlayarak düğümler arası bağlantılarla gösterilen kurallara göre müşterinin risk düzeyi hakkında karar verilmektedir.

Şekil 3: Karar ağacı algoritması



Karar ağacı algoritmasının önemli konularından birisi kuralarda kullanılacak özniteliklerin belirlenmesidir. Sınıf değerinin belirlenmesinde en çok fayda yaratan özniteliğin belirlenerek öncelikle onunla ilgili kuralların koyulması önem arz etmektedir. En iyi bölen özniteliğin bulunması için entropi ve gini endeksi gibi yöntemler kullanılmaktadır. Entropi, sistemdeki rassal bir değişkenin belirsizlik ölçüsü ve bu sistemdeki bilginin değerinin ölçüsüdür. Tüm durumların net olarak bilinmesi durumunda (örneğin her sınıfta eşit veri olması) entropi değeri sıfır olacakken belirsizlik arttıkça bu değer büyüyecektir. Entropi değerlerinin bulunabilmesi için her özniteliğin her kategorisinin gerçekleşme sayıları belirlenmeli buna göre entropi formülü (Denklem 1) uygulanmalıdır.

$$entropi = -p(a|c) * log_2 p(a|c) \quad (1)$$

t toplam ilgili özniteliğe ait kategorilerin toplam gerçekleme sayısını, i ise ilgili kategorideki sınıflara göre belirlenen gözlem sayısını ifade etmektedir. Entropi hesabında ilk olarak her özniteliğin her kategorisinin sınıflara göre toplam gerçekleşmedeki oranı hesaplanır. İkinci aşama olarak bu oranlara göre her sınıfa göre ilgili kategori için entropi hesaplanır. Bu entropi değerlerinin toplanması ile ilgili kategori için entropi değeri elde edilir. Son aşamada ise her kategorinin entropi değeri öznitelikteki oranına göre ağırlıklandırılır ve toplanarak öznitelik entropi değerine erişilmiş olur.

Sınıf özniteliği için hesaplanmış olan entropiden ilgili özniteliğin entropisi çıkarıldığında bilgi kazancı değeri elde edilir ve bu değer ile en iyi bölen özniteliği seçimi yapılabilmektedir. Çünkü bilgi kazancı değeri ilgili özniteliğin kullanılmasının algoritmaya katkısını temsil etmektedir ve en yüksek bilgi kazancına sahip olan özniteliğin seçilmesi gerekmektedir. Böylece en yüksek bilgi kazancına sahip olan öznitelik kök düğüm olur ve yaprak düğümlere ulaşana kadar aynı işlemler ara düğümler içinde gerçekleştirilir.

Tablo 3: Banka müşterisi risk seviyesi veri seti

Yaş	Cinsiyet	Gelir	Banka Müşterisi	Risk
36-55	Erkek	1000-3500	Hayır	Düşük
56-75	Erkek	1000-3500	Hayır	Yüksek
16-35	Erkek	1000-3500	Hayır	Yüksek
36-55	Erkek	3501-7000	Evet	Yüksek
16-35	Kadın	3501-7000	Evet	Yüksek

Bunlara göre karar ağacı algoritmasının entropi yöntemi ile uygulanısı banka müşterilerinin risklerinin belirlenmesi örneğine ait 5 verilik örnek veri seti (Tablo 3) üzerinden Tablo 4'de adımları ile açıklanmaktadır.

Tablo 4: Entropi değerlerinin hesaplanması

Öznitelik	Kategori	1.Adım					2.Adım		3.Adım	4.Adım	5.Adım
		Düşük Sınıfı Gözlem	Yüksek Sınıfı Gözlem	Toplam Gerçek-leşme	Düşük Sınıfı İçin $p(i\|t)$	Yüksek Sınıfı İçin $p(i\|t)$	Düşük Sınıfı İçin Entropi	Yüksek Sınıfı İçin Entropi	Kategori Entropi Değeri	Öznitelik Entropi Değeri	Bilgi kazancı
		a	b	c=a+b	d=a/(a+b)	e=b/(a+b)	f=-d*$\log_2$d	g=-e*$\log_2$e	h=f+g	k=∑c/5*h	m-k
Gelir	1000-3500	1	2	3	0.33	0.66	0.52	0.38	0.91	0.55	0.17
	3501-7000	0	2	2	0	1	0	0	0		
Cinsiyet	Kadın	0	1	1	0	1	0	0	0	0.64	0.08
	Erkek	1	3	4	0.25	0.75	0.5	0.31	0.81		
Yaş	16-35	0	2	2	0	1	0	0	0	0.4	0.32
	36-55	1	1	2	0.5	0.5	0.5	0.5	1		
	56-75	0	1	1	0	1	0	0	0		
Banka Musterisi	Evet	0	2	2	0	1	0	0	0	0.55	0.17
	Hayır	1	2	3	0.33	0.66	0.52	0.38	0.91		
Risk	-	1	4	5	0.2	0.8			m=0.72		

Bu örneğe göre en yüksek bilgi kazancına sahip olan yaş özniteliği, en iyi bölen öznitelik olarak karar ağacının kök düğümü olacaktır.

Karar ağacı algoritmalarının; modelin anlaşılması ve sonuçların yorumlanması açısından diğer veri madenciliği tekniklerine göre daha kolay olması, verilerin yoğun ön işleme sürecinden geçmeden kullanılabilir olması, aynı anda hem kategorik hem de sürekli değişkenler ile analize uygun olması gibi avantajları bulunmaktadır. Öte yandan ezbere öğrenme, yerel optimale takılma gibi dezavantajları da gözlenmiştir (Akpınar, 2014).

Karar ağacı algoritmalarının genel yapısı kök düğümden yaprak düğümlere doğru devamlı daha iyi ayırıcı düğümler bularak, yapraklara ulaşmayı sağlayacak kuralların koyulmasıdır. Bu amaçla çok çeşitli algoritmalar geliştirilmiştir. Sınıflandırma ve tahmin amaçlı geliştirilen bu algoritmalardan bazıları; ID3, C4.5, CHAID, CART gibi işletme, makine öğrenmesi, tıp, psikoloji, bilişim ve daha birçok farklı alanlarda kullanılan algoritmalardır.

1.4.1.2. Naive-Bayes

Naive Bayes sınıflandırma algoritması, Bayes teoremine dayalı, olayların gerçekleşme sıklığının hesaplandığı koşullu olasılık yöntemini kullan istatistiksel bir sınıflandırıcıdır. Önerme sınıflandırmada kullanılacak olan her özniteliğin istatistiksel olarak bağımsız olduğu üzerine kuruludur. Yani bir özniteliğin sınıfta yarattığı etki diğer özniteliklerin var olup olmamasına bağlı değildir.

Naive Bayes algoritması, önceden sınıfı belirli olan verilerin istatistiklerini kullanarak yeni verilerin hangi sınıfa etiketleneceğinin olasılığını hesaplamayı amaçlar. Böylece her özniteliğin sonuca olan etkisinin olasılık değeri de hesaplanabilmektedir. Eğitim veri seti ile her bir öznitelik ve sınıf arasındaki ilişkinin olasılık değeri öğrenilir ve test setindeki sınıf değeri olmayan veriler için bu olasılıklar kullanılarak özniteliklerin bağımsızlığı varsayımı altında sınıflandırma işlemi yapılır.

Algoritmanın işleyişinde öncelikle sınıfların olasılıkları ve özniteliklerin tüm durumları için sınıf değerlerine bağlı koşullu olasılıkları bulunur. Yeni gelen veriler ise ilgili sınıflar için tüm olasılıklar çarpılması ile elde edilen değerlerden hangisi daha yüksekse o sınıfa atanır (Dean, 2014).

Tablo 5: Naive Bayes algoritması örnek veri seti

		Sınıf Değeri		Koşullu Olasılık	
	j	Evet	Hayır	P(evet/j)	P(hayır/j)
Medeni durum	Evli	a	f	a/(a+b+c)	f/(f+g+h)
	Bekâr	b	g	b/(a+b+c)	g/(f+g+h)
	Dul	c	h	c/(a+b+c)	h/(f+g+h)
Eğitim durumu	< Lise	d	ı	d/(d+e)	ı/(ı+k)
	> Lise	e	k	e/(d+e)	k/(ı+k)

Naive Bayes sınıflandırma algoritması için sınıf değerinin evet ya da hayır, özniteliklerin ise medeni durum ve eğitim durumu olduğu ve Tablo 5 'de gösterilen bir örnek veri setinde öncelikle P(evet/j) ve P(hayır/j) sütunlarında gösterildiği şekilde sınıflara bağlı koşullu olasılıklar hesaplanmalıdır. Tabloda gösterilen a, b, c, ... ,k değerleri her özniteliğin ilgili durumunun ilgili sınıfta kaç kez gözlendiğini ifade etmektedir. Bu bilgilere göre yeni gelen;

- medeni durumu "bekâr", eğitim durumu "> lise"

şeklindeki bir verinin hangi sınıfa ait olduğunu belirlemek için her sınıfa ait koşullu olasılık değerleri çarpılarak karşılaştırılmalıdır. Buna göre yeni veri;

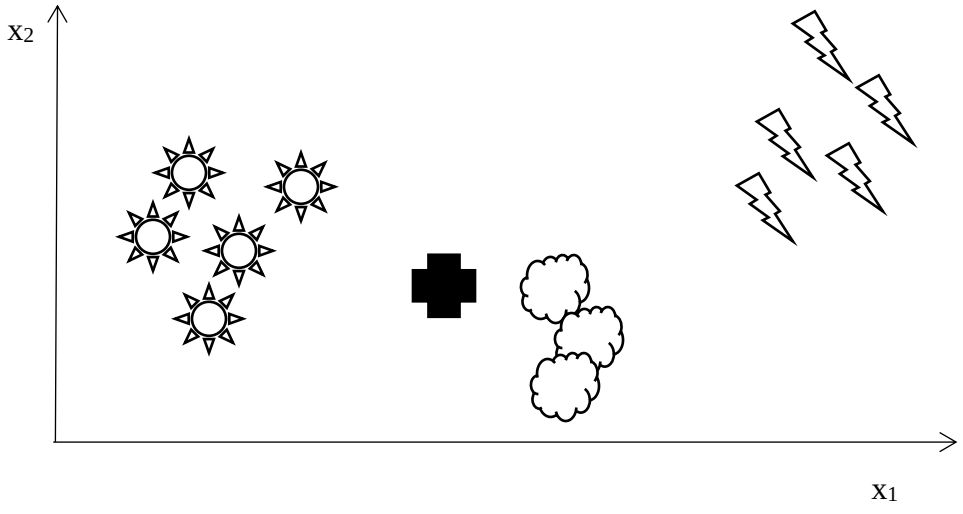
- P(evet/bekar)*P(evet/>Lise)*P(evet) > P(hayır/bekar)*P(hayır/>lise)*P(hayır) ise evet sınıfına
- P(evet/bekar)*P(evet/>Lise))*P(evet) < P(hayır/bekar)*P(hayır/>lise)*P(hayır) ise hayır sınıfına atanır.

Naive Bayes sınıflandırıcılarının, bağımsız olmayan veya büyük hacimli veri setlerinde de hızlı ve başarılı sonuçlar elde edebilmesi, diğer veri madenciliği algoritmalarından uygulama ve anlaşılma bakımından daha kolay olması araştırmacıların ilgisini çekmektedir (Catal ve Nangir, 2017). Ayrıca metin madenciliğinde; metin belgelerinin sınıflandırılması ve duygu analizi çalışmalarında sıklıkla uygulandığı gözlemlenmiştir.

1.4.1.3. K-En Yakın Komşu

Veri madenciliği tekniklerinden uygulama ve çalışma prensibi açısından en basit algoritmalardan biri olan k-en yakın komşuluk, mesafeye dayalı bir sınıflandırma yaklaşımı sunmaktadır. Bu algoritmada veri setindeki tüm verilerin birbirilerine olan uzaklıklarını hesaplanmaktadır. Bu uzaklıklara göre belirlenen komşuluk ilişkileri ile etiketsiz veri, kendisine en yakın komşu ya da komşularının sınıfına atanmaktadır (Akpınar, 2014). Algoritmada önem arz eden hususlardan birisi hangi komşuların sınıfının değerlendirileceğidir. Adından da anlaşıldığı gibi yöntemin başında bir k değeri belirlenmeli ve sınıflandırılacak verinin k adet komşusunun sınıf değerlerinin incelenerek sınıf ataması yapılmalıdır. K değerinin küçük seçilmesi verinin en yakınındaki komşularının sınıfına atanmasına, büyük seçilmesi ise uzağındakilerin de değerlendirilmesi bu nedenle de farklılığı yüksek bir kümeden sınıfının belirlenmesine neden olmaktadır. Bu nedenle algoritmanın etkinliği k değerine göre değişiklik gösterebilmektedir.

Şekil 4: k-en yokun komşuluk algoritması



Şekil 4'ün hava durumunun tahminlenmesi amacıyla kurulan k-en komşuluk modelinin veri setinin sembolize edildiği varsayılsın. Artı şeklinde gösterilen sınıf etiketi olmayan verinin; güneşli, bulutlu ya da yağmurlu sınıflarından birine atanması için k-en yakın komşuluk algoritması kullanılacaktır. Buna göre k değeri 3 seçildiğinde artı nesnesine en

yakın 3 sınıfın bulutlu olması bu verinin sınıf değerinin "bulutlu" olarak atanacağını gösterir. K değerinin 5 olması durumunda kendisine en yakın 5 verinin 3'ünün sınıfı bulutlu 2'sinin ise güneşli olmasından dolayı gene sınıf değeri "bulutlu" olarak atanır. K değerinin 7 olarak belirlendiği durumda ise en yakın olan verilerin sınıflarının 3'ü bulutlu, 4'ü güneşli olacağından "güneşli" sınıfına atanacaktır. Bu nedenle k değerinin belirlenmesi önemli bir konu olarak araştırmacıların karşısına çıkmaktadır. En etkin k değerinin belirlenmesinde, çeşitli k değerleri için algoritmanın tahmin başarı performanslarının değerlendirilesi bir çözüm olabilmektedir. K-en yakın komşuluk algoritması, sağlık alanından, enerji sektörüne, ekonomik çalışmalardan genetik bilimine birçok alanda sınıflandırarak tahmin modellerinde kullanılmaktadır.

1.4.1.4. Yapay Sinir Ağları

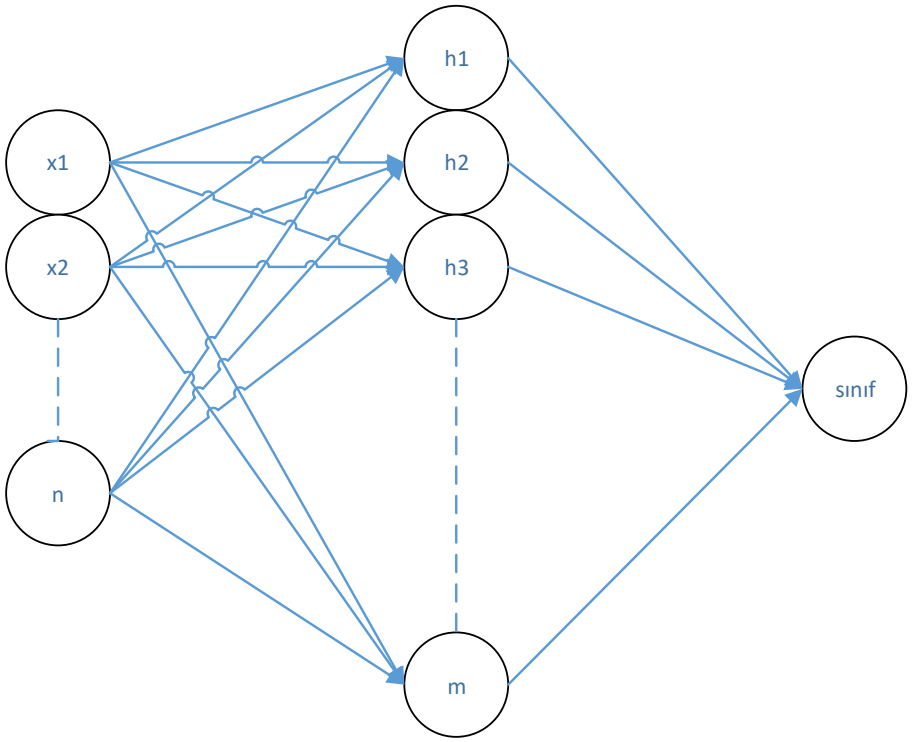
Yapay sinir ağları, insan beyninin öğrenme yapısından esinlenilerek, bu sürecin matematiksel olarak modellenmesi olarak tanımlanabilir. Sinir sisteminin bir parçası olan insan beyni nöron adı verilen sinir hücrelerinden oluşmaktadır. Biyolojik sistemde nöronlar arasındaki bağlantıların gelişmesi ile öğrenme, düşünme ve algılama gibi bilişsel davranışlar ortaya çıkmaktadır (Kalogirou, 2000). Yapay sinir ağlarında da girdi ve çıktı verilerinin nöronlara dönüştürülmesi ve aralarındaki bağlantılar vasıtasıyla eğitilmesi ile öğrenme algoritması ortaya çıkarılmaktadır. Bu sayede öğrenerek yeni bilgi üretmek ve keşfedebilmek gibi yetenekleri otomatik olarak gerçekleştirebilen bir bilgisayar sistemi elde edilmiş olur.

McCulloch ve Pitts'in 1943 yılında insan beyninin hesaplama yeteneğini taklit ederek, elektrik devreleriyle basit bir ağ modellemesi ile yapay sinir ağları literatüre girmiştir. Onları 1949 yılında Hebb, öğrenme teorisi ile ilgilenerek izlemiştir. 1957 yılında ise sıçrama noktası sayılabilecek çalışmayı, Perceptron adı verilen girdilerden bilgileri alarak çıktı üretmeye yarayan sistemi geliştiren Rosentblatt yapmıştır. (Şimşek, 2006) 60'li yılların sonuna doğru perceptron kuramının çözme becerisinin yetersizliği üzerine çıkarılan makaleler ile yapay sinir ağları çalışmalarına olan ilgi azalsa da 80'li yıllarda ortaya konan, geriye yayılım algoritması (Rumelhart ve McCelland, 1986) gibi yeni

çalışmalar ile tekrar popülerliğini kazanmıştır (Akpınar, 2014). Günümüzde halen karmaşık örüntülerin sınıflandırılmasındaki başarısı sayesinde, görüntü işlemeden, hisse senedi seçimine, metin sınıflandırmadan, ekonomik çalışmalara geniş bir yelpazede uygulama alanı bulmaktadır (Emir v.d., 2012).

Basit yapıdaki bir yapay sinir ağında girdi ve çıktı değişkenleri nöronlar şeklinde tanımlanır ve Şekil 5'de gösterildiği gibi giriş, gizli ve çıktı katmanı olmak üzere 3 katman bulunur. Geriye Yayılım Ağları olarak bilinen ağ yapısı katmanlar arasında tam bağlantı olan, ileri beslemeli ve denetimli olarak eğitilen yapay sinir ağılarıdır. Geri besleme ya da bir katmanın atlanması söz konusu olmayan bu ağ yapısında nöronlar vasıtası ile bilgiler bağlantı ağırlıklarının çeşitli işlemlerden geçirilmesi ile giriş katmanından gizli katmana oradan da çıktı katmanına iletir. Elde edilen çıktılar ile gerçek çıktılar karşılaştırılır ve her çıktı nöronu için hata sinyalleri hesaplanır ve bu sinyaller gizli katmandaki ilgili nöronlara aktarılır. Böylece hata sinyallerine göre bağlantı ağırlıkları güncellenir ve eğitim başarısı yüksek bir ağ elde edilmesi sağlanır.

Şekil 5: Yapay sinir ağları algoritması



Geriye yayılım ağları uygulamalarında eğitim sürecine başlanmadan önce, ağın yapısı yani kaç gizli katman olacağı ve bunlarda kaçar adet nöron bulunacağı, öğrenme katsayısının ve momentum değerlerinin belirlenmesi gibi öncül işlemlerin yapılması gerekmektedir. Literatürdeki birçok problem üç katmanlı ağ ile çözülebilmektedir. Fakat daha hızlı öğrenme ve daha başarılı bir öğrenme işlemi için katmanlar eklenmesi gerekebilir (Akpınar, 2014). İdeal bir katman ve katmanların içereceği nöron sayısı bulunmadığından eğitim sırasında deneme yanılma yöntemi ile ilgili veri setine uygun ağ yapısı elde etmek mümkündür. Benzer şekilde ağın öğrenme hızını temsil eden öğrenme oranı ve momentum değerlerinin de deneme yanılma yöntemiyle araştırılması ile en küçük hata değerini verecek ağın parametreleri belirlenebilir.

1.4.1.5. Destek Vektör Makineleri

Vapnik ve Chervonenkis'in (1960) önerdiği öğrenmenin temel teorisinin geliştirilmesi ile 1992 yılında Vapnik, Boser ve Guyon tarafından ortaya çıkarılan "Destek Vektör Makineleri" yöntemi sınıflandırma ve regresyon analizinde kullanılan bir denetimli öğrenme yöntemidir. Diğer sınıflandırma yöntemlerinde olduğu gibi destek vektör makinelerinde de amaç, eğitim setinden öğrenilen bilgiler ışığında çeşitli girdiler ile çıktı değerinin tahmin edilmesini sağlayacak modelin kurulmasıdır.

$$D = \{(x_i, y_i) | x_i \in R^p, y_i \in \{-1,1\}\}_{i=1}^{n} \quad (2)$$

Yukarıdaki gibi bir veri setinde x_i girdi değişkenleri, y_i ise çıktı değişkenini temsil ederken DVM aşağıda formülize edilmiş amacı ve kısıtı sağlayacak optimizasyon problemini çözmeyi hedefler (Boser v.d., 1992; Cortes ve Vapnik, 1995);

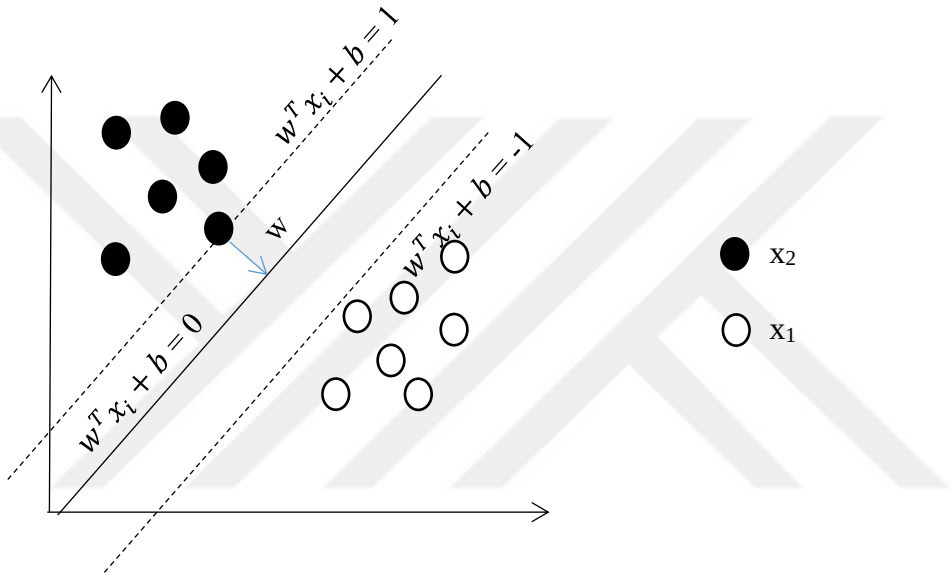
$$\min \quad \frac{1}{2} w^T w \quad (3)$$

$$y_i(w^T x_i + b) \geq 1 \quad (4)$$

DVM bu çok boyutlu uzayda farklı sınıflara ait destek vektörler arasındaki uzaklığı maksimize etmeyi sağlayacak doğrusal bir hiperdüzlem bulmaya çalışır. Amaç fonksiyonunda yer alan "w", Şekil 6'da iki boyutlu veri için gösterildiği gibi sınıf çizgisi

olarak belirlenen bu hiperdüzlem ile destek vektöreler arasındaki uzaklığı temsil etmektedir. Dolayısı ile yapısal riski minimize etme prensibine göre çalışan ve dış bükey optimizasyonuna dayalı bir yöntemdir (Soman, Loganathan ve Ajay, 2011).

Şekil 6: Karar destek makineleri algoritması



Veri setinin çok boyutlu ve karmaşık olması, gürültü içermesi gibi nedenlerden dolayı Denklem 5'de belirtilen en küçükleme modeline ε katsayısı eklenerek belirli bir hata ile ayrılması hedeflenmiştir (Li v.d., 2009). Formülde, hata katsayısı olan ε'nin ceza parametresi C ile gösterilmiştir. Bu sayede belirli bir hata değerinden daha yüksek ε değerine sahip olan verilerin cezalandırılarak önlenmesi hedeflenmiştir.

$$\min \quad \frac{1}{2} w^T w + C \sum_{i=1}^{n} \varepsilon_i \tag{5}$$

$$y_i(w^T x_i + b) \geq 1 - \varepsilon_i \tag{6}$$

$$\varepsilon_i \geq 0 \tag{7}$$

Doğrusal olarak ayrılamayan veya karmaşık veri setleri için en küçükleme probleminin kısıt fonksiyonuna Kernel fonksiyonu eklenmiştir. Φ ile gösterilen kernel fonksiyonu, verilerin daha üst düzey boyutlarda bir uzaya dönüştürülmesi ile kolayca ayrılabilecek duruma getirmeyi hedeflemektedir. (Akpınar, 2014)

$$\min \quad \frac{1}{2} w^T w + C \sum_{i=1}^{n} \varepsilon_i \tag{8}$$

$$y_i(w^T \Phi(x_i) + b) \geq 1 - \varepsilon_i \tag{9}$$

$$\varepsilon_i \geq 0 \tag{10}$$

Literatürde bu amaçla geliştirilen çeşitli kernel tipleri yer almaktadır. Bunlardan en sık kullanılanlar (Akdoğan, 2014);

Lineer; $K\left(x_i, x_j\right) = x_i^T x_j$ (11)

Polinom; $K\left(x_i, x_j\right) = (\gamma x_i^T x_j + r)^d , \gamma > 0$ (12)

Radyal Temelli Foksiyon; $K\left(x_i, x_j\right) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma > 0$ (13)

Sigmoit; $K\left(x_i, x_j\right) = (\gamma x_i^T x_j + r)$ (14)

Kernel fonksiyonlarda parametrelerinin seçilmesinde çeşitli yöntemler kullanılmaktadır. Bunlardan biri ızgara aramasıdır. Bu yöntemde kullanılacak olan parametreler için değişim aralıkları belirlenir ve parametre grupları oluşturulur. En düşük hata değerini veren parametre grubu DVM yönteminde kernel fonksiyonun parametresi olarak kullanılır. Izgara aramasında parametrelerin hangi aralıklarda seçilip aranacağı önemli bir unsurdur. Bu amaçla Hsu ve arkadaşları (2003) üstel olarak artan parametre dizileri oluşturmanın daha iyi sonuç verdiğini vurgulamışlardır. Bu çalışmaya göre öncelikle geniş aralıklı üstel artan parametre dizileri oluşturulmalı ve arama yapılmalı, sonrasında en küçük hata değerini veren parametre değerlerine yakın olan değerler arasında tekrar arama yapılmalıdır. Böylece başta kabaca seçilen alandan, daha iyi bir alan bulunur ve sonrasında daha dar bir alanda arama yapılır.

Destek Vektör Makineleri yöntemi doğrusal olmayan sınıflandırmadaki başarısı, yüksek güvenilirlik oranı ve ezbere öğrenmeye karşı güçlü olması sayesinde regresyon ve sınıflandırma amaçlı birçok çalışmada tercih edilmektedir. Pazarlama, finans vb. işletmenin temel alanlarında uzun veya kısa dönemli tahminleme modelleri, sağlık

alanında hastalık ya da tedavi tahmin modelleri, görüntü işleme modelleri ve doğal dil işleme modelleri destek vektör makinelerinin sıklıkla kullanıldığı alanlarıdır.

1.4.2. Kümeleme

Etiketlenmemiş veri setlerinden bilgi çıkarımında sıklıkla kullanılan konulardan biri olan kümeleme; verilerin aynı kümede yer alacak gözlemler arası benzerlik ve gruplar arası farklılık maksimum olacak şekilde gruplandırılmasıdır. Kümeleme analizinde, küme içi homojenliğin ve kümeler arası heterojenliğin en yüksek düzeyde olabilmesi için gözlemi oluşturan birimler çeşitli özelliklerine göre kümelere ayrılır. Böylece aynı kümede yer alan birimler birbirileri ile benzeşirken, farklı kümelerdeki birimler ile benzeşmezler (Nakip, 2003).

Kümeleme analizi, ekonomi, finans, pazarlama, sağlık vb. birçok alanda karar vericilere yol göstermektedir. Gelişmişlik ya da ekonomik göstergelerine göre benzer ülkeleri gruplanması, finansal performanslarına göre benzer firmaların gruplanması, satın alma alışkanlıklarına göre benzer müşterilerin gruplanması, şikâyet ya da belirtilere göre benzer hastaların/hastalıkların gruplanması gibi çeşitli alanlarda kümeleme analizi algoritmaları kullanılmaktadır (Bramer, 2007).

Gözlemlerin benzerliklerine göre gruplandırılması için yapılacak olan kümeleme analizinin aşamaları şu şekildedir;

1. Problemin belirlenmesi
2. Kullanılacak olan benzerlik ölçüsünün belirlenmesi
3. Kullanılacak olan kümeleme yönteminin ve buna göre kullanılacak olan kümeleme algoritmasının belirlenmesi
4. Küme sayısına karar verilmesi
5. Kümelerin yorumlanması
6. Kümelerin geçerliliğinin test edilmesi.

Gözlem birimlerinin birbiri arasındaki uzaklık/yakınlıklarının veya benzerliklerinin belirlenmesi için benzerlik ölçüsü kullanılmaktadır. Bu ölçü sayesinde elde edilen

benzerlik ya da uzaklık değerlerinin ilgili kümeleme algoritmasında kullanılmasıyla kümeleme analizi yapılmış olur.

Kümeleme analizinin geniş uygulama yelpazesi nedeniyle, benzerliklerin ölçülmesinde ve kümelerin oluşturulmasında farklı ölçü ve yöntemler kullanılmaktadır. Benzerlik ölçüleri gözlem değerlerinin oransal ya da aralıklı ölçekli olması, sayımla elde edilmesi ya da ikili gözlemlerle elde edilmesi gibi faktörlere göre çeşitlilik göstermektedir. Literatürde en sık kullanılan ölçülerden biri Öklid uzaklığıdır. Bunun dışında Ki-Kare Uzaklık Ölçüsü, Phi-Kare Uzaklık Ölçüsü, Kare Öklid, Büyüklük Farkı Yapı Farkı (Pattern Difference), Lance ve Williams farkı gibi benzerlik ölçüleri de kullanılmaktadır.

Veri setine ve çalışmanın amacına göre kümeleme yöntemleri farklılık gösterebilmektedir. Gelişen teknoloji ile bilgisayarların işlem güçlerinin artması ve biriken verinin hacminin büyümesi, öncelerde hiyerarşik ve hiyerarşik olmayan yöntemler olarak kolayca ikiye ayrılabilen kümeleme yöntemlerinin sayısını ve gruplarını arttırmıştır (Dillon ve Goldstein, 1984 & Akpınar, 2014). Buna göre literatürde sıklıkla kullanılan bazı kümeleme yöntemleri şöyledir;

- Hiyerarşik Temelli Kümeleme Algoritmaları; BIRCH, CURE, ROCK, Chameleon
- Bölümleyici Kümeleme Algoritmaları; K-ortalamalar, K-medoids, K-prototypes, K-median, Canopy
- Yoğunluk Temelli Kümeleme Algoritmaları; DBSCAN, OPTICS, DENCLUE
- Izgara Temelli Kümeleme Algoritmaları; STING, CLIQUE
- Alt Uzat Arama Algoritmaları; PROCLUS, SUBCLUE

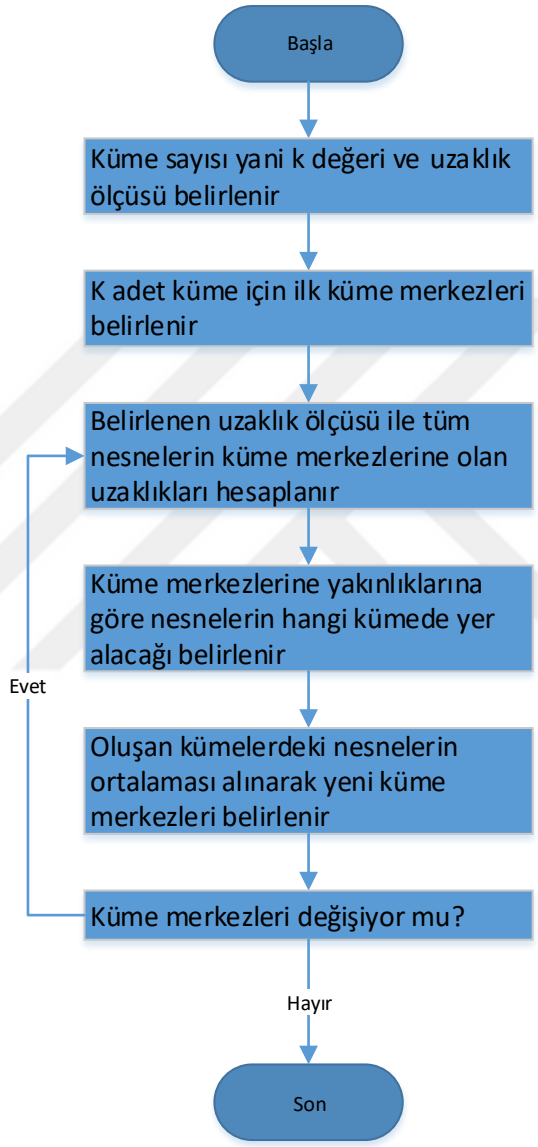
Hiyerarşik yöntemlerde nesnelerin ve özniteliklerin birbirileri ile yakınlık ilişkilerine bağlı olarak oluşturulan ağaç yapısına benzeyen bir yapı oluşturulmaya çalışılır. Bu yapıya göre, başlangıçta her nesne kendi başına bir küme oluştururken ikili birleşmeler ile tüm nesneler tek bir küme oluncaya dek işlem devam ettirilir. Dendogram adı verilen ağaç yapısının gövdeden yapraklara ya da yapraklardan gövdeye doğru istenilen yerden bölünmesi ile kümeler elde edilir. Ağacın bölüneceği yer yani ideal küme sayısı küme içi

uzaklığın minimum ve kümeler arası farklılıkların maksimum olduğu seviyedir. Bu seviyenin belirlenmesinde çeşitli teknikler kullanılmaktadır fakat son karar uygulayıcıya bırakılmıştır. (Hair v.d., 1998). Hiyerarşik kümeleme algoritmalarından bazıları, Merkezi Kümeleme (Centroid) algoritması, Ortalama Bağlantı (Average Linkage) algoritması, Ward algoritması, Birch algortiması ve Chamellon algoritmasıdır (Karypis v.d., 1999).

Bölümleyici yöntemlerde ise nesnelerin kaç kümeye ayrılacağı önceden bilinmektedir. Önceden belirlenen bu küme sayısı algoritmaya parametre olarak verilir. Bu sayede nesneler belirlenen k adet kümeye hangi kümenin ortalamasına daha yakınsa o kümede yer alacak şekilde paylaştırılır. Hiyerarşik yöntemler gibi ikili benzerlikler incelememesinden dolayı daha hızlı çalışan hiyerarşik olmayan yöntemlerde de kaç küme olacağının önceden belirlenmesinin yaratacağı güvensizliğin önlenmesi için algoritma tekrarlanarak optimum sonuca ulaşmak hedeflenir. (Dunham,2006). Bölümleyici kümeleme algoritmalarından bazıları ise, K-ortalamalar algoritması, K-medoids algoritması ve yoğunluk temelli algoritmadır.

Hızı ve uygulama kolaylığı ile büyük veri setlerinde çalışmaya daha uygun olması nedeniyle k-ortalamalar algoritması literatürde sıklıkla kullanılmaktadır (Dean, 2014). 1967 yılında MacQueen tarafından geliştirilen k-ortalamalar algoritmasından, kümeleme özelliğinin yanı sıra veri ön hazırlama ve veri setinin barındırdığı gizli örüntülerin tanımlanması için de faydalanılmaktadır. K- ortalamalar algoritmasının en önemli aşamalarından birisi küme sayısı yani k değerinin belirlenmesidir. Tesadüfi olarak belirlenen k değeri yerine literatürde; uzman görüşü ile ya da örneklem üzerinde hiyerarşik kümeleme algoritmalarının uygulanması, buna göre veri setinin eğiliminin belirlenmesi ile uygun k değeri atanması önerilmektedir (Akpınar, 2014). K- ortalamalar algoritmasının akış şeması Şekil 7'de gösterilmektedir. Algoritma; nesne, küme ve iterasyon sayısı ile orantılı bir karmaşıklığa sahiptir.

Şekil 7: K-ortalamalar algoritması



K-ortalamalar algoritması akışında yer alan küme merkezine uzaklığın hesaplanması kategorik verilerde mümkün olamamaktadır. Bu amaçla uzaklık ölçüsü yerine uyum katsayıları kullanılarak ve küme merkezleri için ortalama alınması yerine modların belirlenmesi çözüm vermektedir (San v.d., 2004). Öte yandan K-ortalamalar yöntemi ile elde edilen kümelerin değerlendirilebilmesi için geliştirilen çeşitli yöntemler vardır. Bunlardan bazıları; Siluet katsayısı, Davies-Bouldin Endeksi ve F ölçütüdür.

İKİNCİ BÖLÜM

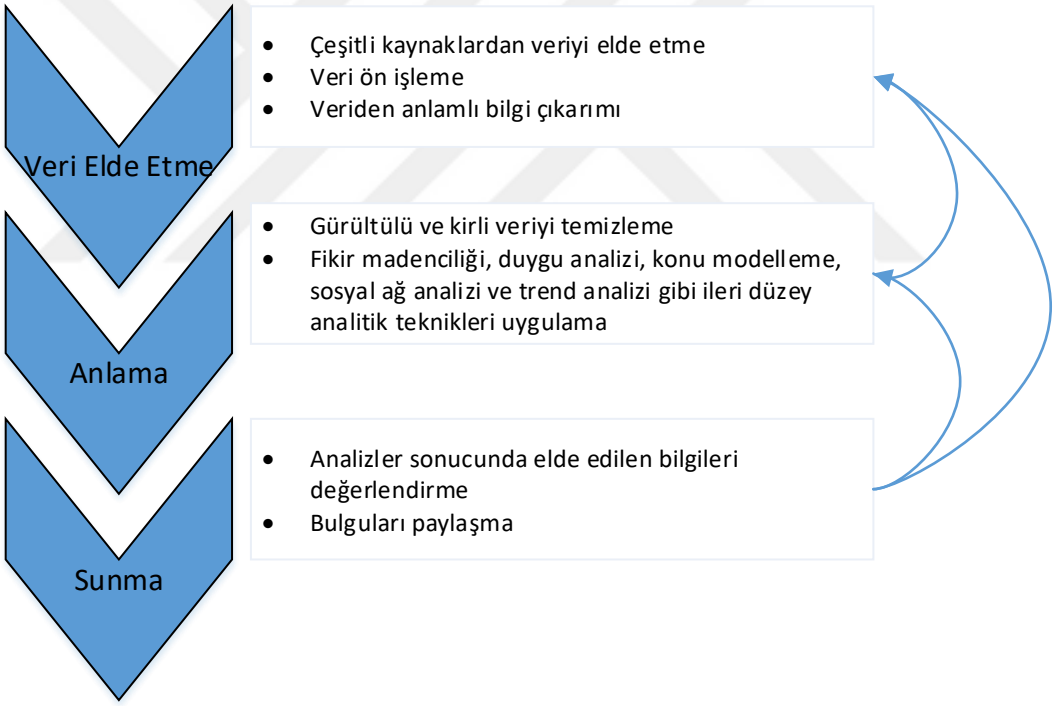
SOSYAL MEDYA ANALİTİĞİ

Kullanıcılarına çevirim içi olarak alışveriş yapabilme, görüşlerini paylaşabilme gibi olanakları sağlayan çeşitli içeriklerin yaratıldığı ve paylaşıldığı internet temelli uygulamalar grubuna sosyal medya adı verilmektedir (Stephen ve Toubia; Kaplan ve Haenlein, 2010). Sosyal medya; Twitter, Facebook gibi sosyal ağların yanı sıra blog, isteyen herkesin fikirlerini yazabildiği çevirim içi sözlükler, haber siteleri ve daha birçok yapısal olmayan veri içeren web kaynağını kapsamaktadır. Bu özelliği ile sosyal bilimlerde veri analitiği ve benzetim modelleri üzerinde çalışan araştırmacılar için önemli bir büyük veri kaynağıdır (Cioffi-Revilla, 2010).

Gelişen teknoloji ile beraber sosyal medya kullanımı da hızla artmaktadır. Her gün milyonlarca kullanıcı sosyal medya aracılığı ile iletişim ve etkileşim kurmaktadır. Örneğin 2014 yılı rakamlarına göre Türkiye'de 35 milyonun üzerinde internet kullanıcısı bulunmaktadır. Bilgi kaynağı olarak sosyal medyanın kullanılması birçok alanda etki yaratabilmektedir. Reklam, tanıtım gibi konularda kullanıcılara ulaşma kolaylığı sağlaması nedeniyle de sosyal medya analitiği çalışmalarına olan ilgi artmıştır. Kişilerin fikir ve görüşlerini paylaşabildikleri iletişim platformu olarak da tanımlanabilen sosyal medya; işletmelere ürünleri ve markaları hakkında mevcut ya da potansiyel müşterilerin görüşlerini kolayca öğrenme fırsatı sunar (Agrawal v.d., 2011). Diğer bir deyişle kişilerin kendi kendilerine düşündükleri şeyleri sosyal medya platformlarında paylaşması ile bu monolog diyaloğa dönüşmektedir (Hansen v.d., 2010). Alışılagelmişin dışında bir hızla iletişim kurmayı sağlaması, birçok marka, sanatçı ve siyasetçiyi sosyal medya kullanmaya itmektedir. Kitle ve algı yönetimi gibi alanlardaki Twitter kullanımının artması Twitter'ın "sosyal bir megafon" olarak tanımlanmasını sağlamaktadır (Baloğlu, 2015). Dolayısı ile sosyal medya kanallarını doğru kullanan kurum ve kişiler olumlu etkilerinden faydalanabilmektedirler. Örneğin Twitter'da bir kullanıcının x ürünü için yaptığı yorum, takipçilerin sayısına, yorumun etkileme gücüne göre hızla yayılıp dakikalar içerisinde binlerce kişinin konuştuğu bir konu haline gelebilmektedir. Bu da iyi yönetilebilirse ilgili

işletme için avantaj sağlayabilecekken, Twitter'ın kullanılmaması ya da bu gibi kullanıcı yorumlarının önemsenmemesi gibi hatalı davranışlar ile dezavantaja dönüşebilmektedir. Her gün milyonlarca içeriğin oluştuğu sosyal medya işletmelere pazarlama alanında önemli ölçüde değer yaratmaktadır. Sosyal medya, işletmelerin geleneksel bir veri toplama aracı olan anket ile ulaşması mümkün olmayan bir kitleden görüş toplamasını sağlar. Elde edilen bu verilerin uygun analizleri ile anlamlı bilgilerin ortaya çıkarılması Şekil 8'de süreci görselleştirilen sosyal medya analitiğinin amaçlarından birisidir.

Şekil 8: Sosyal Medya Analitiği süreci

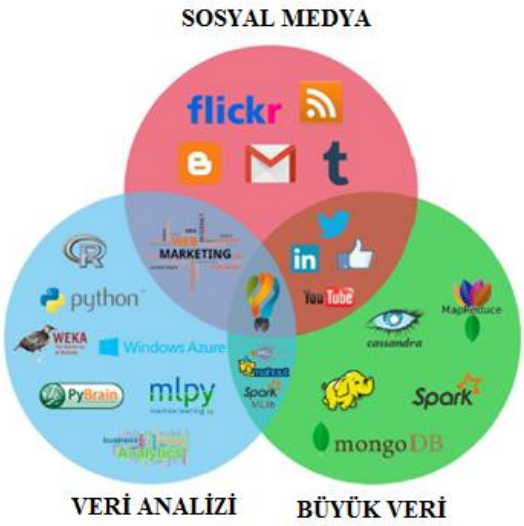


2.1. Sosyal Medya Analitiği: Veri

Sosyal medya verisi büyüklüğü, zenginliği, dinamikliği, insan davranışını temsil başarısı ile bireyleri, grupları ve toplumları anlama imkânı sunmaktadır (Batrinca ve Treleaven, 2015). Araştırmacıların, yeni ve özgün teknikler ile otomatik olarak bu verilerin toplanması, düzenlenmesi ve analiz edilmesi üzerinde yapılan çalışmaları gelişen teknoloji ile her geçen gün artmaktadır. Teknolojik gelişmeler, sosyal medya platformlarında oluşturulan verinin, birikerek büyük hacimli veri yığınlarına dönüşmesine

neden olmaktadır. Eski teknikler ile sadece yığın olarak görülen bu veri setleri gelişen depolama ve analiz teknikleri ile değer kazanmış, anlamlı bilgiler elde edilebilen veri setleri haline gelmiştir. Sosyal medya platformlarında metin, fotoğraf, video, ses, tıklama verisi ve fare hareket verisi gibi çeşitli formatlarda veri üretilmektedir. Geçmiş zamanlı ya da gerçek zamanlı beslemeler şeklinde olabilen bu veri setleri yapısal olmayan ya da yarı yapısal formattadırlar. Bu özellikler sosyal büyük veri kavramının ortaya çıkarmaktadır ve Şekil 9'daki gibi büyük veri, veri analizi ve sosyal medya kavramlarının iç içe geçmiş olduğunu göstermektedir.

Şekil 9: Sosyal büyük verinin yapısı



Kaynak: Bello-Orgaz, 2016

Wamba ve arkadaşlarının 2016 yılında yaptıkları kategorizasyonla, sosyal medya verisini genel hatları ile demografik veri, ürün verisi, psikografik veri, davranış verisi, yönelim verisi ve konum verisi şeklinde kategorilere ayırmaktadır. Bu kategorizasyon sayesinde sosyal medyadan elde edilen veriler, içerdikleri bilgiler ve çıktılarına göre gruplandırılmıştır.

Demografik veri, yaş, cinsiyet, eğitim gibi kullanıcıların kendi profillerinde halka açık olarak paylaştıkları bilgileri kapsamaktadır. (Kaplan ve Haenlein, 2010). Bireyler

hakkında bilgi içeren bu veri seti birikerek işletmeler için anlamlı çıktılar üretilmesinde kullanılabilir. Örneğin, Amazon kullanıcıların demografik özelliklerine göre kişiye özel teklifler sunarak kullanıcıları ile uzun süreli ilişki kurmayı hedeflenmiştir (Nemschoff, 2013). Öte yandan internet bankacılığı kullanan müşterilerin profilleri incelenerek elde edilen verilere çeşitli veri madenciliği yöntemleri uygulanarak, pazarlama stratejileri geliştirilebilmektedir (Gürsoy ve Bilgin, 2016).

Ürün verisi, kişilerin belirli bir ürün ya da marka hakkında sosyal medya üzerinden yaptıkları yorumlardır. Kullanıcıların kendi profillerinde ya da işletmenin resmi profilinde görüş ve önerilerini bildirmesi ile ürün verisi oluşmaktadır. Derlenen ürün verisi sayesinde işletmeler pazarlama stratejileri geliştirmede, müşteri şikâyetlerini takip edip hızla geri dönüş yapmada ve müşteri ihtiyaçlarını öğrenmede avantaj elde etmektedir. Bu amaçla Apple iPhone 5'i piyasaya sunmadan önce ürünün özelliklerini tanıtan reklamları çeşitli iletişim kanalları aracılığı yayınlatmıştır. Bu reklamlara sosyal medya aracılığı ile yapılan yorumlar ve görüşlere göre firma, bireylerin alma arzusunu arttıracak promosyon ve pazarlama stratejilerini belirlemeye çalışmıştır (Moore, 2014). Bu tez çalışmasında kullanılacak veri seti de bir işletme ve ürünleri hakkında sosyal medyada kişilerin yazdığı yorum ve görüşler olduğu için ürün verisi sınıfına girmektedir.

Psikografik veri, sosyal medyadan müşterilerin kişilik, değer, tutum, ilgi alanı ve yaşam şekilleri hakkında bilgi toplanmasına, böylece kişiler hakkında daha detaylı bilgi sahibi olunmasını sağlar. Böylece herkes için tek bir pazarlama stratejisi uygulamak yerine farklı ilgi alanlarına göre farklı stratejiler uygulanabilmektedir. Örneğin sosyal medya üzerinden müşterilerin ilgi duydukları spor dallarının belirlenmesi ile daha çok müşteriye ulaşılacak bir strateji geliştirmek mümkündür.

Davranış verisi; müşterilerin geçmiş satın alma davranışlarının sosyal medya üzerinden elde edilmesidir (Kietzmann v.d.,2011). Bu sayede bir Avustralya bankası olan UBank, müşterisi olmayan kişilerin çalıştıkları bankalar hakkında sosyal medyadan yaptıkları olumsuz yorumları inceleyip, bu kişileri kendi bankalarına çekmeye çalışmışlardır (Schlagwein, 2014).

Yönelim verisi; sosyal medya sayesinde kişilerin gelecek alışverişlerine dair istekleri ve aktiviteleri önceden elde edilebilmektedir. Birçok sosyal medya platformunda karşımıza çıkan "bunları beğenebilirsiniz" temalı reklamlar bu verilerin analizi ile ortaya çıkmaktadır. Kullanıcıların geçmişte ziyaret ettikleri ürün ya da marka profillerinden, yaptıkları yorumlardan yapılan çıkarımlar ile gelecek alışverişler yönelimleri hakkında veri elde edilebilmektedir.

Konum verisi; kullanıcıların gerçek zamanlı olarak sosyal medyadan bildirdikleri konumların derlenmesi ile elde edilir. Bu veri sayesinde bölgesel kampanyalar yapılarak müşteri memnuniyeti sağlanmaya çalışılır. Örneğin Foursquare adlı sosyal medya platformunda yer bildirimi yapılmasının ardından civardaki işletmelerden kampanya haberleri gelmesi bu verinin analiz edilip kullanılması ile mümkün olmaktadır.

Her geçen gün gelişen teknoloji ile artan sosyal medya platformu, oluşan verinin de hacim ve karmaşıklık olarak büyümesine neden olmaktadır. Bu özelliği ile önemli bir büyük veri kaynağı olan sosyal medya, açığa çıkartılmayı bekleyen anlamlı bilgileri sayesinde sektör ve araştırmacılar için daha da cazip duruma gelmektedir. Bunun için gelişmiş analiz algoritmalarından faydalanmak araştırmacılara ve işletmelere kullanım kolaylığı ve daha güvenilir sonuçlarla bilgilerin açığa çıkarılmasını sağlamaktadır.

2.2. Sosyal Medya Analitiği: Araçlar

2.2.1. Veri Toplama Araçları

Sosyal medya analitiğinde açık kaynaklı veri tabanlarından hazır derlenmiş veri setleri, ticari veri sağlayıcılardan ücret karşılığı satın alınan veri setleri veya uygulama programlama ara yüzleri (API) aracılığı ile kullanıcının kendi derlediği veri setleri kullanılabilmektedir.

Açık kaynaklı veri tabanları, araştırmacıların hiçbir ücret ödemeden belli kurum ya da kişilerce derlenmiş veri setlerine ulaşmalarını sağlamaktadır. Wikipedia, UCI, TUIK ve Dünya Bankası veri bankası gibi kaynaklar veri tabanlarına ücretsiz giriş imkânı sunmakta ve araştırmacılara çeşitli alanlarda veri setleri sağlamaktadır. Önemli bir sosyal medya

platformu olan Wikipedia, dumps.wikimedia.org adlı web sayfası ile var olan tüm içeriklere ulaşım imkânı sunmaktadır (Batrinca ve Treleaven, 2015). UCI adlı web sayfasında yaklaşık 360 veri seti paylaşılmaktadır. İlgili veri setleri çeşitli araştırmacılar tarafından toplanıp üzerine çalışmaların yapılmasının ardından diğer araştırmacıların da faydalanabilmesi için paylaşılmaktadır (http://archive.ics.uci.edu/ml/). Bu gibi veri setlerinin elde edilmesinden sonra çeşitli veri analiz teknikleri uygulanarak araştırılan konu hakkında detaylı bilgiye sahip olunabilir. Fakat bu çalışma beceri ve zaman gerektirdiğinden, bazı şirketler online olarak bu hizmeti sağlamaktadır. Topsy.com ve keyhole.co bu amaçla kurulmuş, kullanıcılarına Twitter verilerinin analizi hizmeti sunan uygulamalardır. Benzer şekilde Google Trends uygulaması ile anahtar kelimelerin Google da aratılma istatistiklerini vermektedir. Tek bir anahtar kelime hakkında bilgi verebildiği gibi birden çok konu hakkında da karşılaştırmalı sonuç imkânı sunmaktadır.

Açık kaynaklı veri tabalarının yanı sıra sosyal medya verilerine ticari veri sağlayıcılar aracılığı ile de ulaşmak mümkündür. Birçok sosyal medya platformu uygulama programlama ara yüzleri ile kısıtlı veri derlenmesini sağlamakta ve fazlanı ticari veri sağlayıcıları aracılığı ile satmaktadırlar. Gnip dünyanın en büyük sosyal veri sağlayıcılarından birisidir. Verisini sattığı ilk partneri Twitter'ın ardından portföyüne Tumblr, Foursquare gibi çeşitli sosyal medya platformlarını da eklemiştir. Ham veri sağlamasının dışında kurumlara sosyal medya analitiği çalışmaları da yapılmaktadır. Türkiye'de de Artiwise, Botego, Semanticum gibi firmalar müşterilerine sosyal medya analitiği hizmeti sunmaktadır.

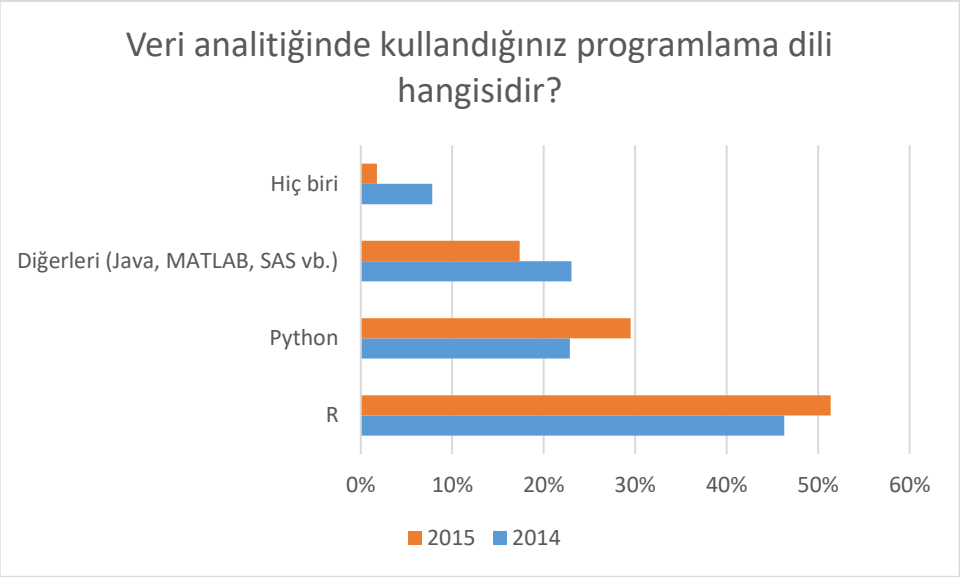
Uygulama programlama ara yüzleri, işletim sistemleri ya da kütüphanelerin diğer programlara sağladığı fonksiyonlar bütünüdür (Ardıç ve Göktürk, 2009). Ayrıca araştırmacıların sosyal medya platformlarının veri tabanlarındaki depolanmış verilere ulaşmasını sağlayan komutlar kümesi olarak da ifade edilebilir. Günümüzde birçok sosyal medya platformu yazılım geliştiricilerin ve araştırmacıların faydalanabilmesi amacıyla bu ara yüzlere sahiptir. Twitter uygulama programlama ara yüzü ile halka açık profilli kullanıcıların tweetlerine ulaşım imkânı sunmaktadır. Bu amaçla https://dev.twitter.com/

web sayfasından kayıt olup R, python ve C gibi çeşitli araçla istenilen anahtar sözcüklerle ilgili tweetleri elde etmek mümkündür. Fakat bu işlemler için 15 dakikada en fazla 180 sorgu gibi limitleri bulunmaktadır (Twitter, 2015). Yapılan sorgular sonucunda, kullanıcı adı, tweet metni, retweet sayısı, favori sayısı, konumu gibi bilgileri içeren Json formatlı dosya oluşturulmaktadır. Bu gibi çeşitli yollarla elde edilebilen sosyal medya verisi analiz edilmediği haliyle sadece yer kaplayan olumsuz bir unsur olarak görülebilir. İşletmeler için çeşitli konularda anlamlı bilgilerin çıkarılabileceği analizlerin uygulanması ile veri yığınlarından, işe yarar bilgi setlerine dönüşüm sağlanabilir.

2.2.2. Analiz Araçları

Sosyal medya verisinin analizinde R, python, MATLAB ve C gibi programlama dilleri ve Weka, Rapidminer gibi yazılımlar kullanılabilmektedir. KDnuggets adlı işletme analitiği, büyük veri ve veri bilimi hakkında güncel araştırmaların yayınlandığı web sayfasında; 2014 ve 2015 yıllarında veri analitiği, veri bilimi ve veri madenciliği alanlarında çalışanlara yapılan bir ankette kullanıcılara hangi programlama dilini daha çok tercih ettikleri sorulmuş ve Şekil 10'da gösterilen sonuçlar elde edilmiştir. R ve python programlama dilleri, veri analitiğinin önemli bir alt dalı olan sosyal medya analitiği için geliştirilen kütüphaneleri ile rapidminer kullanışlı operatörleri ile araştırmacılar tarafından sıklıkla tercih edilen analiz araçlarından bazılarıdır. Bu tez çalışmasında kullanılan bu üç araç hakkında detaylı bilgiler aşağıda verilmektedir.

Şekil 10: Programlama dili tercihleri



Kaynak: http://www.kdnuggets.com/polls/2015/r-vs-python.html

R; istatistik, veri analitiği ve makine öğrenmesinde kullanılabilen açık kaynak kodlu bir programlama dilidir (R Project, 2017). Geniş bir yelpazede istatistiksel analiz ve görselleştirme yeteneği olan program ayrıca geliştirilmeye açık olması ve kullanım kolaylığı gibi özellikleri ile ilgi çekmektedir. Kullanıcılar tarafından geliştirilen paketler/kütüphaneler sayesinde; dosya okuma ve yazma, görselleştirme ve çeşitli çözüm algoritmaları R programına eklenebilmektedir. Veri analitiği konusunda birçok paket içeren R programlama dili sosyal medya analitiği üzerine çalışan araştırmacıların da sıklıkla kullandığı bir programdır. Sosyal medya analitiğinde kullanılmak üzere geliştirilen kütüphanelerden bazıları ve amaçları şöyledir;

- twitterR- Twitter verisi toplama
- tm- metin madenciliği
- wordcloud- kelime bulutu oluşturma
- igraph- ağ analizi oluşturma ve görselleştirme
- RTextTools- otomatik metin sınıflandırma
- topicmodels- konu modelleme

- RCurl- “www” verisi toplama
- XLM- xlm dosyalarını okuma ve yazma

Python, sistem, uygulama ve veri tabanı yazılımı programlama gibi çeşitli alanlarda yazılım geliştirilebilen bir programlama dilidir (Python, 2017). R programlama dilinden farklı olarak python sadece bir istatistiksel analiz veya veri analitiği programı değildir. Fakat içerdiği kütüphaneler, açık kaynak kodlu olması, kullanım kolaylığı ve online kaynakların çokluğu nedeniyle veri bilimcileri tarafından da sıklıkla kullanılmaktadır. Sosyal medya analitiği için geliştirilen python kütüphanelerinden bazıları ve amaçları şu şekildedir;

- Twitter- Twitter verisi toplama
- Tweepy- Twitter verisi toplama
- NLTK- metin ön işleme ve doğal dil işleme
- Pandas- veri analitiği

Açık kaynak kodlu programlama dillerinin yanı sıra Rapidminer da sağladığı çeşitli fonksiyonlarla araştırmacıların veri bilimi ile ilgili her türlü ihtiyaçlarını karşılamayı ve yüksek kaliteli analizler yapabilmelerini hedefleyen bir analitik platformudur (Rapidminer, 2017). Kolay anlaşılır ve uygulanır görsel tasarımı ile araştırmacıların ilgisini çeken rapidminer sosyal medya analitiği alanında da çözümler sunmaktadır. Sosyal medya analitiğinde kullanılabilecek rapidminer operatörlerinden bazıları ve amaçları şöyledir;

- Twitter- Twitter verisi toplama
- Text Processing- Metin verisi ön işleme
- Web mining- web madenciliği
- Aylien- metin madenciliği

2.3. Sosyal Medya Analitiği: Teknikler

Sosyal medya analitiğinin temel noktalarından biri çıktı olarak ne elde edilmek istendiğidir. Veriden anlamlı bilgi çıkarılmasını amaçlarken bu bilgini ne olduğu, hangi

tür veriden hangi tür yöntemler ile elde edileceğine ışık tutar. Bu nedenle araştırmacı analitik sonucunda neyi öğrenmek istiyorsa ona göre veri tipini, veri kaynağını, kullanılacak aracı ve uygulayacağı analiz tekniğini belirlemelidir. Sosyal medya analitiğinde sıklıkla kullanılan veri analiz tekniklerinden bazıları; konu modelleme, duygu analizi (fikir madenciliği), sosyal ağ analizi, eğilim analizi ve müşteri bağlılık analizidir.

Konu modelleme, sosyal medya platformlarından alınan büyük ölçekli metin veri setinden en baskın konuları ayıklamayı hedeflemektedir. Bu sayede kullanıcıların ilgi alanları, önem arz eden konu başlıkları, görüşleri hakkında bilgi edinilebilmektedir. Forumlara yazılan yorumlar, Twitter ya da Facebook'daki durum güncellemeleri ve Foursquare, Tripadviser gibi sitelerdeki kullanıcı yorumları gibi metin verileri konu modelleme tekniğinde kullanılabilmektedir. Bu teknik sayesinde halkın politik görüşlerinde, toplumsal konulardaki fikirlerinde, markalar hakkındaki görüşlerinde yer alan baskın temalar ve konular belirlenebilmektedir. Konu modelleme işletmelere sosyal medyadan elde ettikleri kullanıcı görüşlerini müşterilerin online sesi olarak değerlendirme olanağı da sağlamaktadır (Özdağoğlu v.d.,2016).

Duygu analizi, konu modelleme gibi kullanıcılar tarafından yazılmış metinleri girdi olarak kullanmaktadır. Kişilerin yazdıkları metinlerden bahsettikleri konu hakkındaki duygularının olumlu mu, olumsuz mu yoksa nötr mü olduğunun ya da o anki duygu durumlarının mutlu mu, üzgün mü vb. olduğunun anlaşılmaya çalışılmaktadır. Duygu analizi ile sosyal medya kanallarından kişilerin belirli ürünler ya da markalar hakkındaki görüşleri derlenebilmekte ve bu görüşlerin duyguları belirlenebilmektedir. Benzer şekilde siyasi bir parti, bir sinema filmi, toplumsal bir olay vb. şeklinde kişilerin görüşlerini beyan ettiği herhangi bir konuda, kullanıcıların metinleri yazarken hissettikleri duygular anlaşılabilmektedir. Bu tez çalışmasında duygu analizi tekniği kullanıldığından gelecek bölümde detaylı bir şekilde açıklanacak ve örneklendirilecektir.

Sosyal ağ analizi, sosyal medyadaki ve kullanıcılar arasındaki genel ilişki yapısını, bağlantıları ve kolayca gözlemlenemeyen ilişkileri grafiksel yöntemlerle kolay anlaşılabilir hale getirmek amacıyla kullanılan bir tekniktir. İşletme ve politika alanlarında

gruplar ve toplumlar arası ilişkilerin araştırılmasında sosyal ağ analizinden faydalanılmaktadır (Hanneman ve Riddle, 2005; Hansen v.d., 2010). İnsan hayatının her alanında –aile, iş, sosyal hayat- sosyal ağ yapıları bulunmaktadır. Sosyal ağ analizi ile bu yapıların belirgin özelliklerinin ortaya çıkarılıp, olası değişiklikler karşısında ağ yapısının göstereceği tepkilerin tahmin edilebilmesi geliştirilecek stratejilerin etkinliğinin artmasını sağlayacaktır (Codal ve Coşkun, 2016).

Eğilim analizi, geçmiş verileri kullanarak piyasa eğilimlerinin ve müşteri davranışlarının tahminlenmesini amaçlamaktadır. Eğilim analizi, yapılan kampanyalar hakkında sosyal medyadan toplanan müşteri görüşleri, online olarak yapılan satış rakamları gibi sosyal medyadan elde edilebilen geçmiş olaylara ait verileri analiz ederek politikalar geliştirilmesine yardımcı olmaktadır.

Müşteri bağlılık analizi, sosyal medya aracılığı ile yapılan online aktivitelerin başarısını ölçmeyi, kullanıcıların ilgili ürün, marka vb. ile ilişkisinin durumu hakkında bilgi vermeyi ve devamlılığını sağlayacak süreçler geliştirmeyi hedeflemektedir (Zailskaite-Jakste ve Kuvykaite, 2012). Sosyal medya platformlarının gelişmesi ile ürünler ve müşteriler, halk ve politikacılar gibi alakalı grupların sosyal medya aracılığı ile ilişki kurması ya da var olan ilişkilerini geliştirmesi üzerine yapılan çalışmalar hız kazanmıştır. Bu ilişkilerin incelenmesi ve sonrası için politika geliştirilmesi de müşteri bağlılık analizinin bir parçasıdır. İşletmelerin ve markaların sosyal medya üzerinden yapılacak kampanyalar veya reklam çalışmaları ile müşteri bağlılığını kurabilmeleri ya da koruyabilmeleri müşteri bağlılık analizinin önemli bir alanıdır.

ÜÇÜNCÜ BÖLÜM

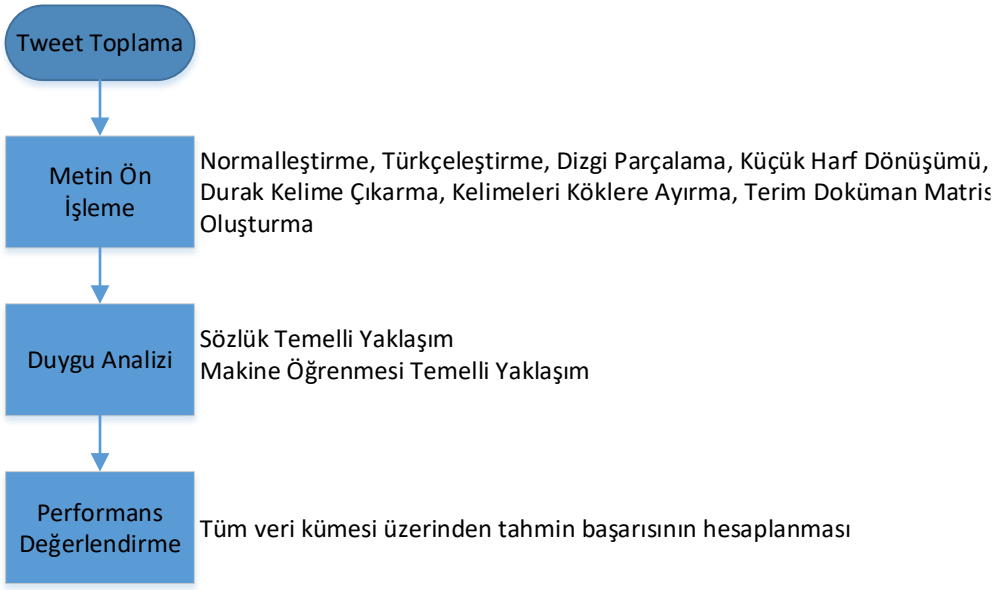
DUYGU ANALİZİ

Metin analizi/madenciliğinin önemli bir alt başlığı olan duygu analizi genel anlamıyla, kişilerin yazdıkları metinlerden görüşlerinin, tutumlarının, yaklaşımlarının veya duygularının belirlenmeye çalışılmasıdır (Pang ve Lee, 2005; Stieglitz v.d., 2014). Artan sosyal medya platformları ile kullanıcıların paylaştıkları sübjektif metinler de artmıştır, bu da ilgili metinlerin analiz edilip anlam çıkarılması şeklinde bir araştırma alanı ortaya çıkarmıştır (Xia v.d, 2011). Duygu analizi literatürde, fikir madenciliği, duygu durum analizi, görüş madenciliği, duygu sınıflandırma gibi farklı isimlerle de yer almaktadır. Duygu analizi teknik yönleriyle mühendislik bilimlerinde, çıkarılan bilgilerin önemiyle de sosyal bilimlerde geniş çalışma alanı bulmaktadır. Mühendislik alanındaki çalışmalar genellikle doğal dil işleme ve daha iyi performanslı tahmin modelleri geliştirmeye yönelikken, sosyal bilimlerde duygu analizi çıktılarının işletmeler, toplum ve ekonomi için anlamlı bulgulara dönüştürülmesi hedeflenmektedir.

Bu alanda yapılan çalışmaların bazıları sadece pozitif ve negatif duyguları belirlemeye çalışırlarken bazıları 0 ile 10 arasında bir skalaya (0-çok negatif, 10-çok pozitif) göre duyguları sınıflandırmışlardır (Pang ve Lee, 2005; Thelwall v.d., 2010). Öte yandan hislerin ve bunların gücünün belirlenmeye çalışıldığı araştırmalar da yapılmaktadır (Neviarouskaya v.d.,2007). Duygu analizi çalışmalarında ihtiyaç duyulan veriler, mikrobloglarda ifade edilen görüşler, ürün/film/otel/restoranlar hakkında yazılan yorumlar, Twitter ve Facebook gibi platformlarda paylaşılan düşünceler, haber makaleleri gibi çeşitli mecralardan elde edilebilmektedir. Sosyal medya platformlarından elde edilen büyük hacimli metin verileri duygu analizi sayesinde, toplumun mutluluk seviyesi, müşterilerin ürünler ve markalar hakkındaki görüşleri, halkın politikacılar veya hükümet hakkındaki görüşleri gibi çeşitli alanlardaki çalışmalarda kullanılan önemli değişkenlere dönüşmektedir (Zhang v.d.,2011; Bollen v.d., 2011; Rahman v.d., 2016). Duygu analizinde kullanılacak metinler; doküman, cümle ve görüş seviyesinde olabilmektedir

(Liu, 2012). Böylece bir metnin tamamını ya da sadece bir cümlenin analizi mümkündür. Kullanıcıların metinleri yazarken belirtmek istedikleri duyguların çıkarılmasını amaçlayan duygu analizinde, kullanılan teknikler genel hatlarıyla sözlük temelli ve makine öğrenmesi temelli olmak üzere iki ayrılmıştır (Medhat v.d., 2014). Makine öğrenmesi temelli analiz teknikleri denetimli öğrenme modelleri kurularak yüksek performans sağlamaktadır. Fakat yüksek model kurma ve etiketleme maliyetleri ile karşı karşıya kalabilmektedir. Sözlük temelli modeller ise uygulaması ve anlaşılması çok kolay olmasına rağmen daha düşük performans ve her yeni çalışmada uzun ön hazırlık süresi gibi dezavantajlara sahiptir(Catal ve Nangir, 2015). Her iki tekniğin de uygulanabilmesi için öncesinde verilerin analize uygun hale getirildiği ön işleme sürecinden geçirilmesi gerekmektedir (Şekil 11). Metinlerdeki anlamsız kelimelerin ve sembollerin temizlenmesi metnin Türkçeleştirilmesi gibi çeşitli adımların izlenmesi ile veri seti uygulama için hazır hale gelebilmektedir.

Şekil 11: Duygu analiz modelinin genel yapısı



3.1. Metin Ön İşleme

Kullanıcıların sosyal medya platformlarında günlük konuşma dili ile ileti girmesi, Twitter iletilerinde 140 karakter sınırlaması olması gibi nedenler metinlerin analiz öncesi

düzenlenmesini gerekli kılmaktadır. Metinlerin analize hazır hale getirilmesi için yapılan ön işlemler aşağıda açıklanmaktadır.

- Normalleştirme; "Çooook güzel" örneğindeki gibi kelimelerin uzatılarak yazılması kelimelerin normalleştirilmesi ihtiyacını doğurmuştur. Öte yandan karakter sınırı nedeniyle izliyorum yerine izlyrm şeklinde yazılan kelimelerin de var olması analizin kalitesini düşürmemek adına önceden tespit edilip düzeltilmesi gereken unsurlardan birisidir.
- Türkçeleştirme; ("ç" , "s" , "ğ" , "ı" ,"ö" , "ü") gibi Türkçe karakterlerin kullanılmaması "çok güzel" yerine "cok guzel" yazılmasına neden olmaktadır ve bunların Türkçe formata getirilmesi gerekmektedir. Böylece çok ve cok kelimesi iki ayrı kelime gibi algılanmayıp tek kelimeye dönüştürülmektedir.
- Dizgi parçalama: bir metinde yer alan tüm kelimelerin ayrı ayrı nesneler haline getirilmesi işlemidir. Böylece "Havalar çok ısındı" cümlesi tek bir cümle olmaktan çıkıp "havalar", "çok", "ısındı" şeklinde üç ayrı özniteliğe dönüştürülmektedir.
- Küçük harf dönüşümü; tüm kelimelerin aynı formatta olmaları için yapılan işlemdir.
- Durak kelimelerin çıkarılması; duygu analizinde ihtiyaç duyulmayan, herhangi bir duygu anlamı içermeyen kelimelerin veri setinden temizlenmesi amacıyla yapılır. Bu sayede veri setindeki nesne sayısının azaltılması hedeflenir. Örneğin ben, sen, o gibi zamirler; ama, fakat, çünkü gibi bağlaçlar durak kelimeler arasında yer almaktadır.
- Kelimeleri köklerine ayırma; nesnelerin boyutunun düşürülmesi ve fiillerdeki zaman ekleri, iyelik ekleri gibi duygu ifade etmeyen unsurların çıkarılması amacıyla yapılan işlemdir.
- Terim doküman matrisi oluşturma; her kelimenin metin ve tüm veri seti içerisinde kullanılma sıklığına göre terim frekans ağırlıklarının belirlenmesi ile oluşturulan matristir. Yani bir kelimenin tüm veri setindeki kelimelere oranla kullanılma sıklığı ile o kelimenin önemi belirlenebilmektedir. Bunun sağlanabilmesi için

terim doküman matris oluşturulurken çeşitli normalleştirme teknikleri kullanılmaktadır (Thelwall v.d., 2010). Bunlar;

- Log frekansı: bulunan frekansların aşağıdaki log fonksiyonu ile normalleştirilmesidir. (wf-ham veri)

$$f(wf) = 1 + \log(wf)\ \forall wf > 0 \quad (15)$$

- İkili frekans: terimin ilgili metinde kullanılıyorsa 1 kullanılmıyorsa 0 değerini alması ile yapılan normalleştirmedir.

$$f(wf) = 1\ \forall wf > 0 \quad (16)$$

- Ters doküman frekansı (IDF): terimlerin metinler içerisindeki etkisinin belirlenmesi için veri setindeki toplam metin sayısının, kelimenin metinde geçme sayısına bölünmesinin logaritması ile elde edilen normalleştirmedir. Frekans ile terimlerin ters doküman frekansları arasında ters orantı vardır.

$$IDF = log\ \frac{veri\ setindeki\ tüm\ metin\ sayısı}{kelimenin\ metinde\ geçme\ sayısı} \quad (17)$$

- Terim frekansı- ters doküman matrisi (TF-IDF): log frekansı ve ters terim frekansı normalleştirmelerinin sentezlenmesi ile oluşturulmuş bir normalleştirme tekniğidir. Terimlerin geçtiği metin sayısının da önem kazandığı bir yöntemdir.

$$TD - IDF = \begin{cases} 0 & if\ wf = 0 \\ 1 + \log(wf) * IDF & if\ wf \geq 1 \end{cases} \quad (18)$$

Metin ön işleme adımları ile Türkçeleştirilmiş, kelimelerin normal formatlarında yazılı olduğu, veri setinde duygu ifade etmeyen kelimelerin yer almadığı metin veri seti oluşturulmuştur. Ardından her kelimenin ayrı bir nesne olarak düzenlendiği, en üst satırında kelimelerin, sol baş sütununda ise metin bilgilerinin yer aldığı, içerisinde de her kelimenin tüm veri seti içerisinde geçme sıklığının yer aldığı matrisin oluşturulması ile analize hazır veri seti oluşturulmuştur.

3.2. Sözlük Temelli Yaklaşımlar

Sözlük temelli duygu analizi yaklaşımlarında kelimeler ya da kelime gruplarının (n-grams) içerdikleri duyguya göre (pozitif/negatif) sınıflandırılması ile duygu sözlükleri oluşturulmaktadır. Bu sözlüklere göre metinlerin duygu durumları belirlenebilmektedir. Her metnin ne kadar pozitif ve negatif sözlük içerdiği incelenip, yüksek sayıda olan sınıfa atanması ile duygusu ortaya çıkarılmaktadır. Buna göre incelenecek olan bir metin daha çok pozitif terim içeriyorsa duygusu pozitif, daha çok negatif terim içeriyorsa negatif ikisi eşit ya da hiçbirini içermiyorsa nötr olarak sınıflandırılmaktadır.

Duygu sözlüklerinin oluşturulmasında manuel olarak duygu sözlüğü oluşturma, yazılı sözlüklerden yararlanarak duygu sözlüğü oluşturma ve bütünce temelli duygu sözlüğü oluşturma gibi yöntemler kullanılmaktadır (Liu, 2010).

3.3. Makine Öğrenmesi Temelli Yaklaşımlar

Makine öğrenmesi temelli yaklaşımlar veri setindeki tüm kelimelerin ayrı ayrı birer öznitelik olarak değerlendirildiği kelime çantası (Bag of Words) tekniği ile sınıflandırma algoritmalarının kullanılmasıdır. Her kelimenin birer özellik olarak Karar Ağacı, Destek Vektör Makineleri, Yapay Sinir Ağları ve Naive Bayes gibi sınıflandırma algoritmalarında girdi olarak kullanılması ile metinlerin sınıflandırılarak duygularının tahmin edilmesi hedeflenmektedir.

Duyguların tahmin edilebilmesi amacıyla kurulan sınıflandırma modellerinde girdi olarak kullanılacak özniteliklerin belirlenmesi gerekmektedir. Bu amaçla metinler ön işlemeden geçirilerek işlevsiz / duygu ifade etmeyen kelimeler ve sembollerin çıkarılması sağlanır. Temizlenmiş veri setindeki her kelimenin bir özniteliği temsil ettiği ve kelimelerin geçme sıklıklarının yer aldığı terim doküman matrisleri oluşturulur. Metin sayısı arttıkça terim doküman matrisinin boyutu da büyüyecektir. Bu nedenle Ki-kare bilgi kazanımı, terim frekansı ya da terim frekansı-ters metin frekansı değerleri kullanılarak temsil gücü yüksek özniteliklerin seçilmesi mümkündür (Zhang, 2013). Ardından denetimli öğrenme modeli için veri setinin bir kısmının etiketlenmesi ile eğitim seti oluşturulur ve sınıflandırma algoritmaları uygulanarak metinlerin duygu durumları belirlenebilir.

DÖRDÜNCÜ BÖLÜM

LİTERATÜR ARAŞTIRMASI

4.1. Literatür Çalışması

Bu bölümde, önceki bölümlerde detaylı bir şekilde açıklanan büyük veri, işletme analitiği, sosyal medya analitiği ve duygu analizi üzerine yapılan çalışmalardan bahsedilecektir.

Xiang ve arkadaşları (2015) bu çalışmada büyük veri analizinin, otelcilikle ilgili önemli konular olan müşterilerin deneyimleri ve tatminleri hakkında faydalı bilgi sağlayıp sağlamayacağını araştırmışlardır. Bu doğrultuda Expedia.com adlı web sayfasından konaklama yapan kişilerin ilgili oteller hakkındaki yorumları ve beğeni oranları elde edilmiştir. Verilerin ön işlemeden geçirilmesinin ardından geliştirilen sınıflandırma modeli ile metin analizi yapılmış ve müşterilerin deneyimleri ile tatmin oranlarının ilişkili olduğunu ortaya çıkarmışlardır. Öte yandan yazarlar bu çalışmanın otelcilik literatüründeki çalışmalara yeni bir soluk getirebileceğinin ve gelecek çalışmalara açık bir alan olmasının üzerinde durmuşlardır.

Hashem ve arkadaşları 2015 yılında yaptıkları çalışmada büyük verinin bulut bilişim alanındaki gelişmeleri konu almışlardır. Bu amaçla iç içe geçmiş olan bu iki kavramı 5 vaka çalışması ile açıklamışlardır. Öte yandan çalışmalarında büyük veri depolama sitemleri ve Hadoop teknolojisi hakkında bilgi verilmektedir.

Büyük veri hakkında genel bilgi verdikleri çalışmalarında Gandomi ve Haider (2015) özellikle analiz yöntemler üzerinde durmuşlardır. Büyük verinin %95'ini oluşturduğu varsayılan yapısal olmayan verilerin analiz yöntemleri detaylı bir şekilde açıklanmıştır. Çalışmada metin analizi, ses analizi, video analizi, sosyal medya analizi ve tahmin edici analizler hakkında geniş bilgi yer almaktadır.

Hmeidi ve arkadaşları (2014) Arapça yazılmış metinlerde metin sınıflandırması üzerine bir çalışma yapmışlardır. Çalışmada Arapça metinlerin sınıflandırılabilmesi için 5 farklı algoritma denenmiş ve karşılaştırılmıştır. Öte yandan çeşitli Arapça ayıklama

tekniklerinin de ilgili sınıflandırma algoritmalarındaki etkinliği incelenmiştir. Bu amaçla WEKA ve RapidMiner yazılımlarını kullanan yazarlar bu çalışmada DVM sınıflandırıcısının daha yüksek doğruluk verdiği sonucuna varmışlardır.

Elragal 2014 yılında Kurumsal Kaynak Planlaması sistemleri ile büyük verinin ilişkisini inceleyen bir çalışma yapmıştır. Bu çalışmada Kurumsal Kaynak Planlaması sistemlerinin büyük veri ile bir arada kullanılabilirliğini araştırmıştır. Yazar büyük verinin genellikle sosyal medya, şeylerin interneti vb. ile ilişkilendirilen bir kavram olduğunu ve Kurumsal Kaynak Planlaması sistemleri ile entegre edilmesinin pek tartışılmadığını bu nedenle çalışmasında bunu temel aldığını belirtmiştir.

Du ve arkadaşları (2014), Çin'deki emlak firmalarının pazarlama stratejileri geliştirmeleri için büyük verinin kullanılmasını önermişlerdir. Bu doğrultuda olası büyük veri uygulamaları, ortaya çıkan sorunlar ve çözümleri ve emlak sektöründe elde edilebilecek rekabet avantajı üzerinde durulmuştur. Çalışmada gayrimenkul pazarlamada büyük veri uygulamalarının gayrimenkul satışlarını artmasını sağladığı sonucuna varılmıştır.

Young (2015), HIV virüsünden kaynaklı hastalıkların oluşumu ve önlenmesinde büyük veriden faydalanılan bir çalışma yapmıştır. Çalışmada sosyal medya ve mobil teknolojilerden yararlanılarak HIV salgınının belirlenmesi için geliştirilen araçlardan bahsetmektedir. Öte yandan büyük veri ile gelişen biyoinformatik, dijital salgın belirleme ve hastalık modelleme ile bunların HIV salgınını tespit etme ve önlemedeki yeri üzerinde durulmaktadır. Çalışmada büyük verinin çeşitli alanlarda değer yarattığı vurgulanmıştır.

Weichselbroun ve arkadaşları (2014) semantik bilgiyi genişletmek için fikir madenciliği çalışması yapmışlardır. Bu amaçla Amazon ve IMBD'den derlenen veriler ile duygu analizi yapılmıştır. Çalışma belirsiz duygu içeren terimleri belirleyip, eğitim kütüphanelerinden elde edilen bağlam bilgisini sağlama ve bu bağlamsal bilgiyi WordNet gibi yapısal bilgi kaynaklarına yerleştirme adımlarını içermektedir. Çalışmanın sonucunda bu bağlamsal ilişkinin istatistiksel doğruluk ve kesinlik üzerinde pozitif bir etkiye sahip olduğu ve bu değerlerde artış meydana geldiği ortaya çıkarılmıştır.

Jifa ve Lingling (2014), veri, veri-bilgi-malumat-irfan (DIKW) hiyerarşisi, büyük veri ve veri bilimi arasındaki ilişkiyi tartışmışlardır. Yazarlar büyük veri çalışmalarında çoğunlukla büyüklük, çeşitlilik ve hız kavramları üzerinde yoğunlaşıldığına ama pratikte esas rol oynayan şeyin değer olduğuna bu nedenle veri bilimi, malumat ve irfanın önemine vurgu yapmışlardır.

Akademik çalışmaların yanı sıra büyük veri literatürüne önemli katkı yapan önemli firma çalışmaları da yer almaktadır. Bunların başında Google şirketinin geliştirdiği "Google Flu Trends" çalışması gelmektedir. Çalışma Google arama motorunda yapılan aramalara göre dünya genelinde grip trend analizi yapılmasını sağlamıştır (Google Flu Trends, 2015). Facebook, Youtube gibi firmalar da kullanıcılarına öneriler yaparken büyük veriyi kullanmaktadırlar. Facebook adlı sosyal paylaşım sitesi reklam ve arkadaş önerisi yapabilmektedir. Benzer şekilde Youtube da kullanıcılarına beğenebilecekleri videoları önermektedirler. Bu çalışmaların ortak özelliği kullanıcıların görüntüleme geçmişinden faydalanarak bir sonraki beğenecekleri/ilgilenecekleri reklam veya videoları kullanıcıya sunmaktır.

Amazon ve eBay gibi çevirim içi alışveriş siteleri de büyük veri uygulamalarından yararlanmaktadırlar. Müşterileri bir ürün satın alırken, o anki ya da geçmiş alışveriş bilgilerinden yararlanarak müşterilere beğenebilecekleri/satın alabilecekleri ürünler öneren sistemleri sayesinde satış rakamlarını yükseltmişlerdir (Chen ve ark., 2012).

Yemeksepeti de büyük veri analizinden yararlanan firmalardan biridir. Firma müşterilerin geçmiş sipariş bilgilerine göre müşterilere özel promosyonlar geliştirmektedir. Öte yandan televizyon programları, futbol karşılaşmaları ve hava durumu gibi unsurların sipariş sayılarına etkilerini aralarındaki ilişkiyi analiz eden çalışmalar yapmaktadırlar (http://hbrturkiye.com/video/vaka-ii-yemek-sepeti-ve-buyuk-veri).

Büyük veri gerçek uygulama denince akla ilk gelen firmalardan biri de Netflix'tir. Netflix'in düzenlediği ödül yarışmalarında film reytinglerini tahminleyici algoritmalar yarıştırılmaktadır. Akademik ve endüstri alanından birçok katılımcının yer aldığı yarışma çeşitli tavsiye edici sistem gelişmesini sağlamıştır (Netflix Prize

(http://www.netflixprize.com//community/viewtopic.php? id=1537; accessed July 9, 2012).

Yukarıdaki çalışmalarda da bahsedildiği gibi büyük verinin birçok çalışma alanı bulunmaktadır bu kısımda tez çalışmasının üzerinde durduğu sosyal medya analizi ve duygu analizi konularında yapılan çalışmalar hakkında bilgi verilecektir.

Genel hatlarıyla sosyal medya analizinin önemi vurgulayan yazarlardan Isson ve Harriott (2013) çalışmasında sosyal medya kişilerin tercihleri, düşünceleri, aktiviteleri, konuları ve hissiyatları hakkında rahat ulaşılabilir veriler sunduğunu vurgulamıştır. Bu verileri analiz etmenin firmalara ve markalarına büyük katkılar sağlayacağını belirtmiştir. Ayrıca Fan ve Gordon (2014) çalışmasında sosyal medya analitiğini gücünü anlatmıştır. 2012 yılın 3800 pazarlamacıya yapılan anketten çıkan 3 temel sorunun; sosyal medya yatırımının dönüşlerinin nasıl olacağı, sosyal medya kullanıcılarının nasıl tanımlanacağı ve bu kullanıcılara yönelik etkin stratejilerin neler olacağı şeklinde belirlendiğini ve sosyal medya analizi kavramının bu sorunları kapsadığını belirtmiştir. Bekmamedova ve Shanks (2014) ise kurumların sosyal medya analizi ile nasıl değer yarabilecekleri hakkında bilgi verdikleri çalışmalarında bir finansal kuruluş için de uygulama yapmışlardır. Çalışma ilgili finansal kuruluşun sosyal medya kanallarını kullanarak geliştirdikleri yaratıcı ve yenilikçi pazarlama kampanyaları ile rakiplerinden ayrıldığı belirtilmiştir.

Çeşitli sosyal medya kanalları üzerinden derlenen veriler ile çeşitli algoritmalar kullanılarak yapılan birçok duygu analizi çalışması bulunmaktadır. Duygu analizi de disiplinler arası bir konu olduğu için hem teknik yani sınıflandırma algoritmalarının geliştirilmesi gibi çalışmalar, hem de analiz çıktılarının kullanılabildiği sosyal bilimler ve sağlık alanlarında yapılan çalışmalar karşımıza çıkmaktadır. Meral ve Diri (2014) ile Çoban ve arkadaşları (2015) Twitter'dan derledikleri Türkçe tweetler üzerinde duygu analizi çalışması yapmışlardır. Türkçe'nin sondan eklemeli bir dil olması nedeniyle duygu analizinde dikkat edilmesi gereken konulara vurgu yapan çalışmalar çeşitli algoritmaların sınıflandırma performanslarını karşılaştırmışlardır. Ahkter ve Soria (2010) ise sıklıkla üzerinde çalışılan Twitter'dan derlenen verilere duygu analizi çalışmalarına alternatif

olarak Facebook profillerinden veri toplayıp duygu analizi yapmışlardır. Diğerlerinden farklı olarak Gunawardena ve arkadaşları (2013) içerikten duygu analizi yapma çalışmalarının benzerini Instagram adlı sosyal paylaşım sitesinden derledikleri etiketler ile yapmışlardır. Kang ve Park (2014) da çalışmalarında Apple Store'dan çevirim içi alışveriş yapan kişilerin bıraktıkları yorumlara duygu analizi yapmışlardır. Duygu analizi sonucu ortaya çıkan özelliklerin Vikor karar algoritması ile değerlendirilmesi ile hangi servislerden memnun kalındığı belirlenmiştir. Jang ve arkadaşları (2013) Youtube'daki yorumlar üzerine yaptıkları çalışmalarında duygu analizi ve izlediği videolar ve yorumlarından yola çıkarak kullanıcı analizi yapmışlardır. Bu sayede yorum yapan kullanıcıların ilgilendikleri alanlar hakkında fikir sahibi olunabildiğini ve bunun firmalar için büyük değer yarattığını belirtmişlerdir. He ve arkadaşları (2013) bir sektörden belirledikleri üç firmanın Facebook ve Twitter üzerinden derledikleri veriler ile müşteri memnuniyetini belirlemeye çalışmışlardır. Bu amaçla takipçi, yorum ve paylaşım sayıları gibi sayısal değerlerin yanı sıra iletiler de metin madenciliği teknikleri kullanılarak değerlendirilmiştir. Mostafa (2013), birçok firma için Twitter paylaşımlarında ilgili firma adlarının yanında en sık kullanılan kelimeleri belirlemeye çalışmıştır. Bu amaçla hava yolu, telekomünikasyon ve sağlık sektörü gibi çeşitli sektörlerden firmaları analiz etmiştir. Bu gibi kullanıcıların firmalar hakkındaki düşüncelerin incelendiği çalışmaların yanı sıra Zheng ve diğerleri (2013) ve Eirinaki ve diğerleri (2012) e-ticaret sitelerinden belirli ürünler için derledikleri metinleri inceleyerek ürünler hakkında yapılan yorumların etkisini ve verilen puanlarla ilgisini araştırmışlardır.

Sosyal medyanın kullanımının finans alanında sıklıkla kullanıldığı konuların başında borsa veya piyasa tahminlemesi gelmektedir. Zhang ve arkadaşları (2011) günlük olarak tweetlerin korku mu umut mu içerdiklerini ölçüp bu değerlerin borsa göstergeleri ile korelasyonlarını analiz etmişlerdir. Buna göre ertesi gün için borsanın yönünü tahminleme imkânı olan bir model geliştirmişlerdir. Bollen ve arkadaşları (2011) ise Twitter'dan derledikleri verileri ruh hali kategorilerine göre sınıflandırıp bunların DJIA kapanış fiyatlarının tahmininde kullanılmasına yönelik bir çalışma yapmışlardır. Günlük kapanış fiyatlarının artış azalış değişimlerinin tahmininde %86,7 oranında doğruluk elde

etmişlerdir. Ardından Mittal ve Goel (2012) de Twitter üzerinden halkın ruh hali skorunu belirleyip bunu borsadaki hareketin tahmininde kullanmıştır. Bu amaçla kullandıkları algoritma %75,56 doğruluk oranını sahiptir ve karşılaştırma yaptıkları Bollen ve arkadaşlarının çalışmasından daha düşük bir oran elde etmişlerdir. Benzer şekilde Bouktif ve Awad (2013) da Twitter'dan derledikleri veriler ile hisse senetleri kapanış fiyatları tahminlemeye çalışmışlardır. Geçmiş çalışmalar ile kıyaslayabilmek için yazarlar çalışmalarında karınca kolonisi optimizasyon algoritması kullanmışlardır. Porshnev ve arkadaşları(2013) da borsa tahminlerinin gücünü arttırabilmek için Twitter mesajlarının duygu analizini kullandıkları çalışmalarında 755 milyon tweet kullanmışlardır. Destek vektör makinaları ve sinir ağları algoritmaları ile DJIA ve S&P500 hakkında tahmin yapmaya çalışmışlardır. Öte yandan Türkiye'de yapılan çalışmalardan biri olan Eliaçık ve Erdoğan (2013)'ın çalışmasında geliştirdikleri yeni bir duygu analiz yöntemi ile finans topluluklarının Twitter'dan topladıkları verisini analiz etmişlerdir. Elde edilen duygu polarite değerleri ile Borsa İstanbul'un haftalık değerleri arasındaki korelasyon incelenmiştir. Bunlara ek olarak Corea (2015) yatırımcıların duygularının borsayı etkileyip etkilemediğini inceleyen çalışmasında yatırımcıların Twitter'a yazdıkları metinlere duygu analizi yapmıştır. Fakat tahminlemede duyguların değil atılan tweet sayılarının daha etkili olduğu sonucuna varmıştır.

Sosyal medya analizinin geniş kullanım alanı bulduğu diğer bir konu da pazarlamadır. Twitter'ın ağızdan ağıza pazarlama için önemli olduğunu vurgulayan çalışmalarında Jansen ve arkadaşları (2009) belirli ürünler ve markalar için bunu analiz etmişlerdir. Ghiassi (2013) ise Twitter'dan bir marka hakkında müşteri yorumlarını derlemiş ve buna literatürdeki modellerden farklı olarak geliştirdikleri algoritmalar ile duygu analizi yapmışlardır. He, Zha ve Li (2013) de 3 pizza markası hakkında Twitter ve Facebook'dan derledikleri tweetlerle metin madenciliği yapmışlardır. Çalışmanın sonucunda firma değerlerinin belirlenmesinde sosyal medya rekabet analizinin önemi ve metin madenciliğinin gücü ortaya çıkarılmıştır. İlgili firmalar hakkındaki iletilerin de metin madenciliği yöntemi ile analiz edilmesiyle hangi markanın hangi sosyal medya kanalında daha çok öne çıktığı da ortaya çıkarılmıştır. Öte yandan Cvijikj ve Michahelles (2013)

geleneksel pazarlama anlayışına yeni bir bakış açısı getirmeyi hedefledikleri sosyal medya pazarlaması konulu bir çalışma yapmışlardır. Bu amaçla firmanın yöneldiği sosyal medya kanalı, kullandığı içerik, içeriğin yayınlama zamanı, markanın Facebook sayfasındaki beğeni, paylaşım ve yorum sayısının pazarlamaya etkisini ölçmeyi amaçlamışlardır. Trattner ve Kappe (2013) ise reklamı görecek kişilerin sayısını ve yatırım geri dönüşünü hızlandıracak olan Facebook reklamı üzerine çalışmışlardır. Bu amaçla gerçek zamanlı ölçümler ile en çok hareket halinde olan kullanıcıların tespitini de analiz etmişlerdir.

Sinema filmlerinin gişe getirilerinin ya da izlenmelerinin tahmini de sıklıkla üzerinde çalışılan bir konudru. Mishne ve Glace (2006) filmler hakkında bloglarda yazılan metinlere duygu analizi yaparak bunun gerçek sinema gişesi rakamları ile korelasyonlarını incelemiştir. Sharda ve Delen (2006) ise filmleri "batanlar" ve "gişe rekorları kıranlar" aralığında kategorilere ayırmak için çalışmışlardır. Bu amaçla geliştirdikleri sınıflandırma problemini sinir ağları algoritması yardımı ile çözmüşlerdir. Diğerlerinden farklı bir sosyal medya kanalı kullanan Asur ve Huberman (2010) Twitter'dan belirli bir konu hakkında veri derlenip geliştirilen modelle piyasa için tahminlemeler yapılabileceğini öngörmüşlerdir. Bu amaçla geliştirdikleri doğrusal regresyon modeli ile Twitterdan sinema filmleri hakkında derledikleri veriler ile gişe kazancı değerlerinin arasındaki ilişkiyi ölçmüş ve gelecek tahmini yapmışlardır. Ayrıca tweetlere duygu analizi yaparak ilgili modeli geliştirmeye çalışmışlardır. Joshi ve arkadaşları (2010) ise metin ve metaveri formatında topladıkları veri ile doğrusal regresyon yöntemi kullanarak filmlerin kazançlarını tahminleme üzerine çalışma yapışlardır. Rui ve diğerleri (2013), Twitter gibi sosyal medya kanalları aracılığı ile yapılan ağızdan ağıza iletişimin ürün satışları üzerinde etkiye sahip olduğunu ve bunun yönetilerek avantaja çevrilebileceğini, filmlerin gişe rakamları üzerindeki etkisi üzerinden ifade etmişlerdir. Kim ve diğerleri (2015), Twitter ve Facebook sosyal medya sitelerinden filmler hakkında yapılan yorum sayılarını, haftalık eğilimlerini ve kaç salonda yayınlandığı gibi filmin yayınlanmasına ilişkin girdileri kullanarak gişe rakamını tahminlemeyi hedeflenmişlerdir. Ding ve diğerleri (2016), filmlerin yayınlanmadan önce sosyal medya kanallarında aldıkları beğenilerin gişe rakamına etkisini incelemişler ve pozitif bir etki olduğunu ortaya koymuşlardır. Hur ve

diğerleri (2016), filmler hakkında yapılan yorumlara duygu analizi yaparak gişe rakamlarını tahmin etmeye çalışmışlardır. Bu amaçla çeşitli veri madenciliği teknikleri kullanmışlardır. Bunların yanı sıra sosyal medyanın kullanılmadığı ama filmin yayınlandığı tarih, oyuncuları, devam filmi olup olmaması, hitap ettiği kitle gibi filmin kendine has özelliklerinden yararlanılarak tahmin modeli geliştirilen çalışmalar da yer almaktadır (Zhang, 2009; Ghiassi, 2015). İlgili çalışmalarda çeşitli veri madenciliği tekniklerinden yararlanılarak filmlerin kazançları ya da başarıları tahmin edilmiştir.

Fisher ve Miller (2011) çalışmalarında hükümetlerin sosyal medya analizi kullanarak elde edecekleri avantajlardan bahsetmiştir. Sosyal medya analitiğinin demografik bilgiler, olumlu veya olumsuz düşünülen konular gibi halkın genel fikirlerini ortaya çıkardığı vurgulanmıştır. Benzer şekilde halk üzerine yapılan bir diğer çalışmada Mislove ve arkadaşları (2011) Twitter aracılığı ile Amerikada yaşayanların %1'i oranında kullanıcıya ulaşmış ve bu kişilerin demografik özelliklerini çıkarmışlardır. Twitter dan elde edilen popülasyon ile Amerika da yaşan popülasyonu coğrafya, cinsiyet ve ırk/etnik kökenine göre karşılaştırmıştır. Çalışmanın sonucunda Twitter'dan elde edilen popülasyonun düzensiz olduğu sonucuna varmışlardır. Dodds ve Danforth ise 2011, 2012 ve 2013 yıllarında çeşitli yazarlarla birlikte Tweet'lerden mutluluk ölçümü ve buna bağlı olarak çıkarımlar yapmışlardır. Sosyal medyanın nüfus çalışmalarındaki önemini vurgulayıp, mutluluk haritalarından, obezite oranlarında artışa kadar geniş bir sonuç yelpazesi ortaya çıkarmışlardır.

Sağlık alanında da sosyal medya anlizinden yararlanılarak yapılan pek çok çalışma bulunmaktadır. Culotta (2010) salgınların izlenmesi ve tahminlenmesi için Twitter'ı kullanmıştır. Twitter'dan grip ile ilgili mesajları derleyen yazar, çeşitli algoritmalar kullanarak bu mesajların "Hastalık Kontrol Merkezi" istatistik değerleri ile ilişkisini incelemiştir. Aramaki (2011) 'de benzer bir amaçla grip ile ilgili tweetlerin pozitif ya da negatif olarak sınıflandırılmasında çeşitli makine öğrenmesi tekniklerinin performanslarını karşılaştırmıştır. Bodnar ve Salathe (2013) ise grip ile ilişkili sayılabilecek en az bir kelime geçen tweetler üzerinde regresyon çalışması yapmıştır.

Ayrıca gazetecilik, deprem tahmini, kitap satışlarının belirlenmesi, atıf sayısının belirlenmesi gibi çok çeşitli konularda ve alanlarda sosyal medya analizinden faydalanılarak anlamlı bilgiler elde edilebileceğini gösteren birçok çalışma bulunmaktadır. Gruhl ve arkadaşları(2005) bloglardan derledikleri bilgiler ile kitap satışlarını tahminlemeye çalışmışlardır. Liu (2007) ise MySpace adlı sosyal ağdan kişilerin lezzet hakkındaki görüşlerini toplayıp buradan kişiler hakkında bilgi elde etmeye çalışmıştır. Zhao ve arkadaşları (2011) Twitterdaki içerikler ile geleneksel gazeteyi karşılaştıran bir çalışma yapmışlardır. Twitter'daki konular ile gazetedeki konuları karşılaştırabilmek için çeşitli metin madenciliği yöntemleri kullanılmıştır. Sakaki ve arkadaşları (2010) ise gerçek zamanlı tweetler ile deprem tespiti yaptıkları ile mevcut deprem ölçüm sistemi kadar iyi çalışan bir model geliştirmişlerdir. Öte yandan kullanıcılara anında bilgi e postası paylaşmaları mevcut deprem ölçümü yapan kurumdan öne geçmelerini sağlamıştır. Akademik alanda yaptığı çalışmasında Eysenbach (2011), yüksek atıf alacak çalışmaların, makalenin yayınlandığı ilk üç gündeki tweetler sayesinde tahminlenebileceğini belirtmiştir. Ayrıca çalışmasında "twimpact" değerinin oluşturulmasını önermiştir. Silvia ve arkadaşları (2013) ise fikir madenciliğinin sadece metin analizi ile olmayacağını bunun için diğer türlerdeki verilerin de analiz edilmesi gerektiğini savunmuşlardır. Bu amaçla çalışmalarında Instagram'dan derledikleri etiket verisi ve fotoğraf verisini birlikte analiz etmişlerdir.

Son olarak bu tez çalışmasının da konusu olan sosyal medya üzerinden televizyon reytinglerinin belirlenmesi üzerinde yapılan çalışmalara bakıldığında araştırmacıların özellikle Twitter'a yoğunlaştığı görülmektedir. Wakamıya, Lee ve Sumiya (2011) çalışmalarında Twitter'ın televizyon programları için izleyicinin fikirlerini anlık olarak takip imkânı vermesi üzerine yoğunlaşmışlardır. Bu amaçla atılan tweetlere göre TV izlenme oranlarını tahmin edecek bir sistem geliştirmişlerdir. Yazarlar sistemin avantajları olarak gerçek zamanlıya yakın bir izleme sağlaması, neredeyse sıfır maliyete sahip olması ve büyük bir kalabalığın fikirlerini araştırma şansı vermesini vurgulamışlardır. Aynı yazarlar yine 2011 yılında tweetler ile TV programları arasındaki ilişkiyi inceleyen bir çalışma yapmışlardır. Şimşek ve Özdemir (2012), çalışmalarında Twitter'dan paylaşılan

iletiler ile borsa değerleri arasındaki ilişkiyi araştırmışlardır. Bu amaçla mutlu ve mutsuz sözlükleri oluştururak duygu analizi yapmışlar ve oluşturdukları ortalama mutluluk skoru ile borsa değerlerini karşılaştırmışlardır. Diri ve arkadaşları (2013, 2016) ise seçilen Türkçe TV programları hakkında Twitter'dan veri derlemişlerdir. Bu metinlere çeşitli algoritmalar ile duygu analizi yaparak belirli bir skor atanmış ve bu skorlar ile gerçek reytingler karşılaştırılmıştır. Diri ve arkadaşlarının 2013 yılında yaptıkları çalışmada Twitter da paylaşılan mesajların 140 karakter ile sınırlı olmasından dolayı kişilerin kısaltma kullanarak yazdığı kelimelerin düzeltilmesinin gerekliliğine vurgu yapmışlardır. Öte yandan diğer problemler olarak, tweet atan kesimin genellikle genç ve orta yaşlı olması, her program için aynı sayıda tweet atılmamış olması ve bazı sosyoekonomik sınıfların Twitter'da temsil edilmemesini belirtmişlerdir. Çalışmanın sonucunda sadece tweet sayılarının tek başına reyting belirlemede yeterli olmadığını bunun için duygu analizi sonuçlarının da kullanılarak geliştirilen bir reyting sıralaması yapıldığını ve bu sıralama ile gerçek reyting sıralamasının birbirine yakın olduğunu vurgulamışlardır. Akarsu ve Diri'nin 2016 yılında yaptıkları çalışmada ise Twitter'dan çekilen verilerin temizlenmesi ve düzenlenmesi için makine öğrenmesi algoritmalarının kullanılması üzerine yoğunlaşmışlardır. Çeşitli algoritmaları çeşitli program türlerinin reytinglerinin tahminlemesinde kullanarak performanslarını karşılaştırmışlardır. Pittman ve Tefertiller (2015) da belirledikleri dört program için ikinci ekran aktivitelerini incelemişlerdir. Günümüzde insanların dizi, film vb. programları izlerken akıllı telefon, tablet gibi ikinci ekranlar da kullandığını vurgulayan çalışma, programların Twitter da konuşulması ile izlenme oranlarını karşılaştırmıştır. Benzer şekilde Cheng ve arkadaşları (2016) da çalışmalarında Tayvan da yayınlanan üç TV programının Facebook sayfalarındaki beğeni, paylaşım ve yorumları toplayıp bunlar ile TV reytingleri arasındaki ilişkiyi araştırmışlardır. Oh ve Yergeau (2017), çalışmalarında kurdukları regresyon modeli ile Facebook da ilgili TV programları hakkında yapılan yorum ve beğeniler ile reyting değerleri arasındaki ilişkiyi incelemişlerdir.

BEŞİNCİ BÖLÜM

UYGULAMA

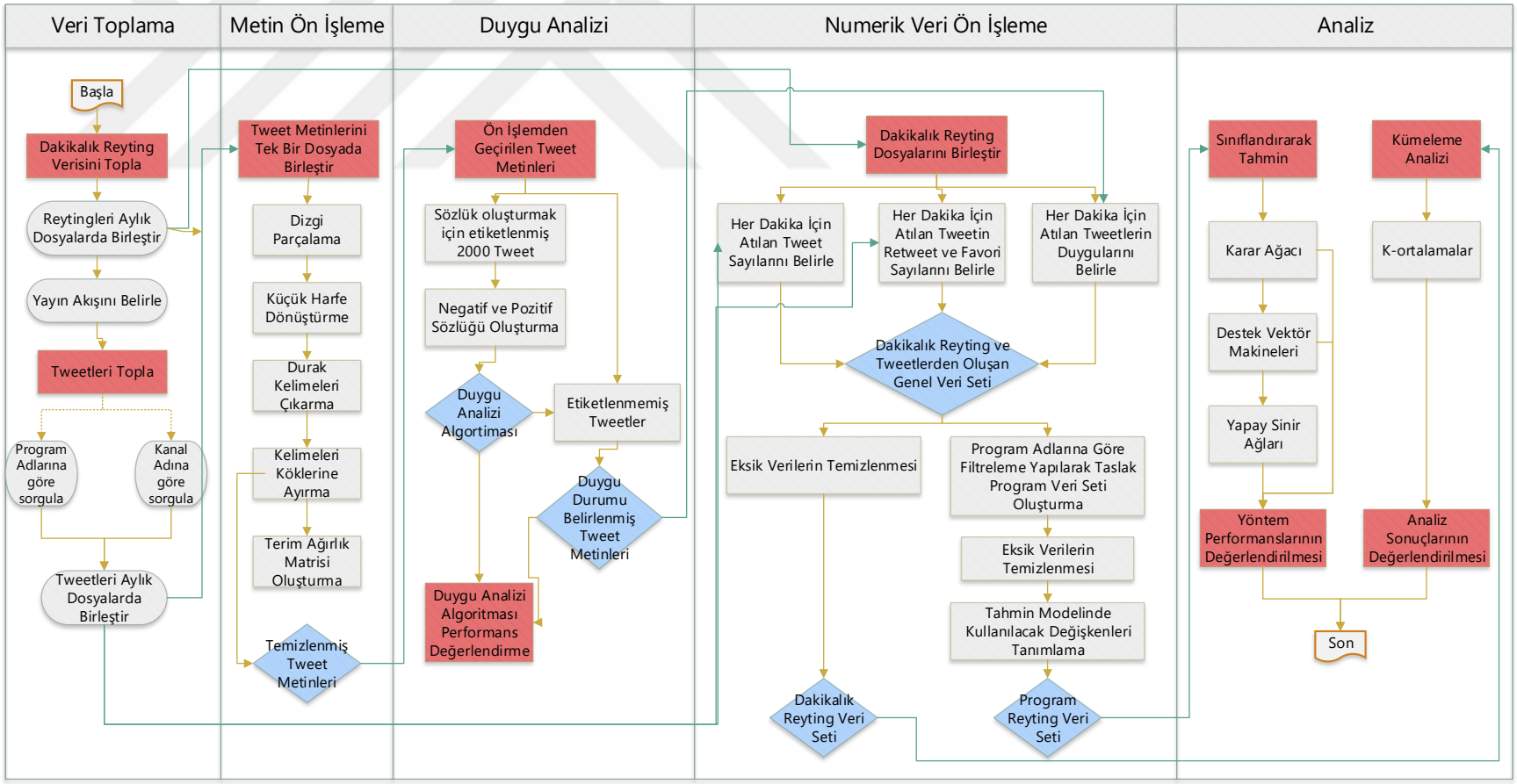
Bu tez çalışmasında bir işletmenin sahip olduğu ürünlerin özellikleri ve kullanıcıların ürünler hakkında sosyal medya üzerinden paylaştıkları görüşlerin nitelik ve nicelik bilgileri ile satış rakamlarının tahmin edilebileceği bir tahmin modeli oluşturulması hedeflenmiştir. Günümüzde sosyal medya aracılığı ile kullanıcı görüşlerine hızlı bir şekilde ulaşmak mümkündür. Bu amaçla dakikada 1 milyondan fazla görüşün paylaşıldığı sosyal paylaşım sitesi Twitter seçilmiştir. Twitter, mikroblog kategorisinde kullanıcılarının 140 karakter kısıtı ile görüşlerini paylaşabildikleri bir platform olarak Temmuz 2006 yılında erişime açılmıştır. "Tweet" adı verilen metinler ile kullanıcılar istedikleri konulardaki fikirlerini diğer kullanıcılar ile paylaşabilmektedir. Hitap ettiği kitleye daha kolay ulaşabilmesi ve yayılma hızının yüksek olması nedeniyle birçok işletme müşterileri ile iletişimde Twitter gibi sosyal medya kanallarını kullanmaktadır. Reklam ve kampanya çalışmalarının Twitter aracılığı ile yapılmasının yanı sıra müşterilerin geri bildirimlerinin de elde edilebilmesi Twitter'ı interaktif bir iletişim aracı haline getirmektedir. Müşterilerin ürünleri hakkında yazdıkları tweetlerin, içerdikleri duygular bakımından analiz edilmesi yani negatif, pozitif veya nötr duygular içerip içermediğinin incelenmesi, işletmelere yeni ürün tasarlama ve pazarlama stratejisi geliştirme vb. konularda karar desteği sağlayabilmektedir. Bu sayede müşterilerin ürünleri hakkındaki yorumları ve önerilerini hızlı bir şekilde elde edebilen işletmeler, gelecek çalışmalarında bunları göz önünde bulundurarak artı değer yaratabilir.

Çalışmada bir işletmenin ürünlerinin özellikleri ve bu ürünler hakkında Twitter'dan yapılan yorumlara duygu analizi yapılarak elde edilen duygu değerleri ile satış rakamlarının tahmin edilmesi ve yeni piyasaya sunulan bir ürünün gelecek dönemde tutulup tutulmayacağının belirlenmesi üzerinde durulmuştur. Bu amaçla işletme olarak bir televizyon kanalı seçilmiş ve işletmenin ürünleri olarak kanalda yayınlanan programlar satış rakamları olarak da reyting değerleri incelenmiştir. Kanaldan Kasım 2015-Haziran

2016 aralığını kapsayan 8 aylık yayın akışı ve dakikalık reyting bilgileri elde edilmiştir. Buna göre 150 farklı program için toplam 340.000 dakikalık reyting bilgisi derlenmiştir. Her program için yayınlandığı aralıkta atılan tweetler incelenmiş ve duygu analizi yardımı ile her tweetin duygusu ortaya çıkarılmıştır. Programlar bahsedilen dakikalık reyting ve tweet değerlerine göre kümelenmiş ve uzman görüşü ile belirlenen reyting sınıfları ile ilgili kümeler karşılaştırılmıştır. Aralarındaki ilişkinin anlamlı çıkması ile 8 aylık süreçte yayınlanan 3075 programa ait özellikler (yayın aralığı, hitap ettiği kitle, ünlü kişilerin olup olmaması vb.), yayın süresince atılan tweet sayıları ve bu tweetlerin duygularının yer aldığı bir tahmin modeli kurulmuştur. Bu tahmin modeli ile bir programın gelecek yayında elde edeceği reyting sınıfının belirlenmesi hedeflenmiştir.

Çalışmanın genel akışı Şekil 12'de gösterilmektedir ve aşağıda aşamalar detaylı bir şekilde açıklanmaktadır.

Şekil 12: Çalışmanın genel akışı



5.1. Veri Toplama

Çalışmanın veri toplama aşaması iki ana parçadan oluşmaktadır. Birincisi kanalın yayın akışı ve dakikalık reyting değerlerinin derlenmesi, ikincisi ise tweetlerin çekilmesidir. Türkiye'de geleneksel reyting ölçümü belirli hanelerdeki televizyonlara konulan reyting cihazları ile yapılmaktadır. 1989 yılında AGB Neilsen firması tarafından yapılmaya başlanan reyting ölçümleri günümüzde TNS firmasınca yaklaşık 2500 haneye konuşmuş cihazla ile yapılmaktadır. Reyting ölçüm çalışmaları temelde 4 basamaktan oluşmaktadır; veri tabanının araştırması, panel seçimi ve yerleştirilmesi, panel ölçümü, ham verilerin işlenmesi raporlanması. Ölçüm cihazlarının konulabileceği haneler ile veri tabanı oluşturulmakta ve veri tabanı araştırması yapılarak izleyici kitlesinin sosyoekonomik statüleri belirlenmektedir. Sosyoekonomik statü grupları oluşturulurken göz önünde bulundurulan hususlardan bazıları aile bireylerine ait eğitim durumu, meslek, gelir seviyesi bilgileridir. SES gruplarının dağılımında bir değişiklik olmaması adına, herhangi bir ailenin ölçümden çıkması durumunda veri tabanında yer alması alan benzer SES grubunda yer alan aileler ile görüşülmektedir (Eyüboğlu, 2012). Şuan A, B, C1, C2, D ve E olmak üzere 6 farklı SES grubu vardır (Ek 1). A grubu yüksek eğitimli ve kalifiye işgücü olarak çalışan (Avukat, doktor vb.) kişileri temsil ederken, E grubu düşük eğitimli, işsiz ya da emekli kişileri temsil etmektedir. Buna göre ilgili hanelere reyting cihazı yerleştirilmesi ile aile bireylerinden hangisinin hangi kanalı, ne zaman ve ne kadar süre ile izlediği kayıt altına alınmaktadır. Bu kayıtlar ham veri olarak her gece 2:00 ile 6:00 arası çekilmekte ve her kanal için analiz edilerek kanallara ait dakikalık "reyting" ve "share" değerleri elde edilip ertesi sabah kanallar ile paylaşılmaktadır. "Reyting" değeri toplam izleme oranını yani ortalama izleyici sayısını, "share" değeri ise ilgili dakikada ilgili programın toplam izlenmeden aldığı payı ifade etmektedir. Kamuya açıklanan reyting(Rtg%) ve share değerleri genellikle "5+", "AB" ve "20+ABC1" sosyo ekonomik statü gruplarına göre oluşturulmuş verilerdir.

Çalışmada incelenecek olan kanalın önceden belirtilen dönemdeki 241 güne ait dakikalık reytinglerini içeren dokümanlar alınmıştır. Örnek dakikalık reyting tablosu Tablo 6'da gösterilmektedir. Her ayın birinci günü 00:00 dakikasından son günü 23:59 dakikasına

kadar var olan tüm reyting bilgileri birleştirilerek aylık dosyalar oluşturulmuştur. Oluşturulan bu aylık dosyalar ile günlük haftalık ve aylık yayın akışları oluşturulmuş ve ilgili aylarda kaç farklı program yayınlandığı belirlenmiştir. Bu bilgiler ışığında hangi aralıklarda hangi programlar için tweetlerin toplanacağı kararı verilmiş ve tweet çekme aşamasına geçilmiştir.

Tablo 6: Dakikalık program ve reyting tablosu örneği

1 OCAK 2016 Cuma Dakikalık
Datalar

Units >>	Share			Rtg%		
Title	Individuals 5+	Individuals SES AB	Individuals 20+ ABC1	Individuals 5+	Individuals SES AB	Individuals 20+ ABC1
A programı	10.19	8.00	11.05	1.44	1.01	1.74
<<02:00 >>	11.70	10.78	12.40	1.80	1.48	2.08
<<02:01 >>	11.56	10.71	11.59	1.76	1.45	1.92
<<02:02 >>	11.17	10.33	11.73	1.70	1.40	1.95
<<02:03 >>	10.84	7.60	11.50	1.63	1.00	1.89
<<02:04 >>	11.09	7.89	12.25	1.64	1.04	1.98

API birçok uygulamanın veri tabanlarında depolanan verinin paylaşılması amacıyla geliştirilen açılımı "Application Programming Interface" olan ara yüz anlamına gelmektedir. Twitter da API aracılığı ile tweetlerin metin içeriği, kullanıcı adı, tarih ve saat bilgisi, retweet ve favori sayısı atıldığı konum ve içerdiği fotoğraf link vb. bilgileri sunmaktadır. Böylece Twitter API'sine ulaşılarak istenen anahtar sözcükleri içeren ya da belirli bir kullanıcı tarafından atılmış tweetlerin elde edilmesi mümkündür. "Search API", "Stream API" ve "REST API" şeklinde 3 farklı API türüne sahip olan Twitter verileri paylaşırken; dakikada 60 tweet çekebilme ya da bir haftadan eski tweetleri görüntüleyememe gibi bazı limitler koymaktadır. Limitlere takılmadan büyük tweet verisi toplayabilmek geliştirilen bazı kodlar ile mümkün olabilmektedir. Bu amaçla çalışmada büyük hacimli bir veri seti oluşturabilmek için Jefferson Henrique tarafından geliştirilen "GetOldTweets"(GOT) projesi kullanılmıştır. Bu proje Twitter'ın API kısıtlarını baypas ederek bir haftadan eski bile olsa istenilen tarih aralığındaki tweetlere ulaşılmasını

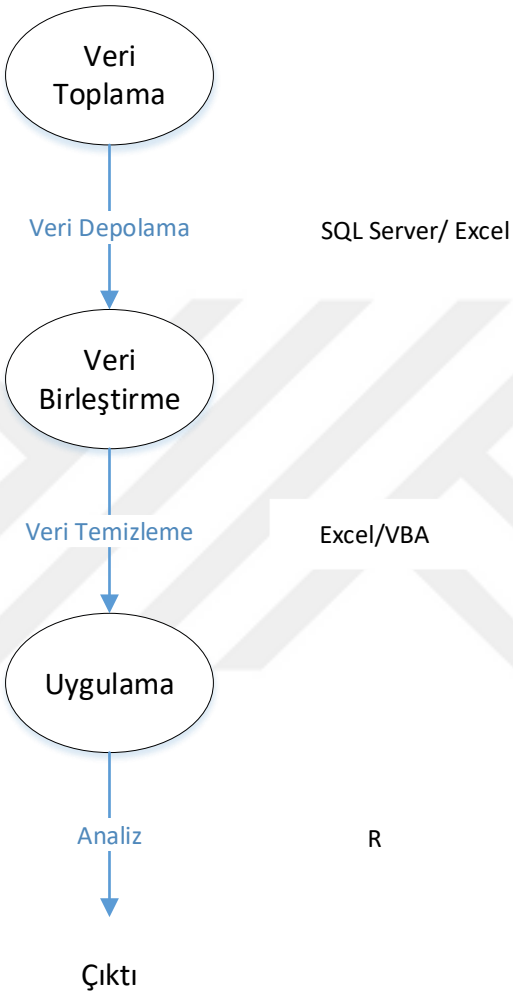
sağlamaktadır. Bunu Twitter'ın arama çubuğuna aranan anahtar sözcükleri yazıp sayfayı otomatik olarak aşağı çekmeye yarayacak bir dizi komut ile gerçekleştiren yazar, platform olarak pyhton kullanmıştır. İlgili proje istenilen tarih aralığında kullanıcı adına ve anahtar kelimeye göre herkese açık profiller içinde arama yapılmasına olanak sağlamaktadır. Arama çıktısı olarak da kullanıcı adı, tarih, tweet metni, retweet sayısı, favori edilme sayısı ve konum bilgilerini sunmaktadır. GOT projesi yazarı tarafından github adlı yazılımcıların projelerini paylaştıkları web sayfasından paylaşılmıştır. İlgili kod grubu web sayfasından indirildikten sonra ara yüz kullanmadan arama dosyası python IDLE ile düzenlenir (aranan anahtar sözcük ve tarihler girilir) ve sonuçlar csv dosyası formatında kaydedilir.

Çalışmada incelenecek olan her televizyon programının hangi tarih ve saatler arasında yayınlandığı bilgisi yayın akışı ile elde edilmişti. Bu bilgilere göre GOT'a aşağıdaki formatta sorgu kodları girilmiştir.

- "python exporter.py --querysearch "a programı lang:tr" --since 2016-05-31 --until 2016-07-01"
- "python exporter.py --querysearch "x kanalı lang:tr" --since 2016-05-31 --until 2016-07-01"

Yukarıda örnekleri verilen kodlar ile ilgili kanal için her ay bir tane olmak üzere toplam 8, televizyon programları için ise 1500'den fazla sorgu yapılmıştır. Bu sorgular sonucunda yaklaşık 1.200.000 adet tweet elde edilmiştir. Elde edilen tweetler x dakikada a programına atılan tweet sayısının belirlenmesi aşamasında kullanım kolaylığı sağlaması adına program adlarına göre etiketlenerek aylık dosyalarda birleştirilmiştir. Genel hatları ile veri toplama ve uygulama süreci Şekil 13'de gösterilmektedir.

Şekil 13: Veri toplama ve uygulama süreci ve kullanılan araçlar



5.2. Metin Verileri Ön İşleme

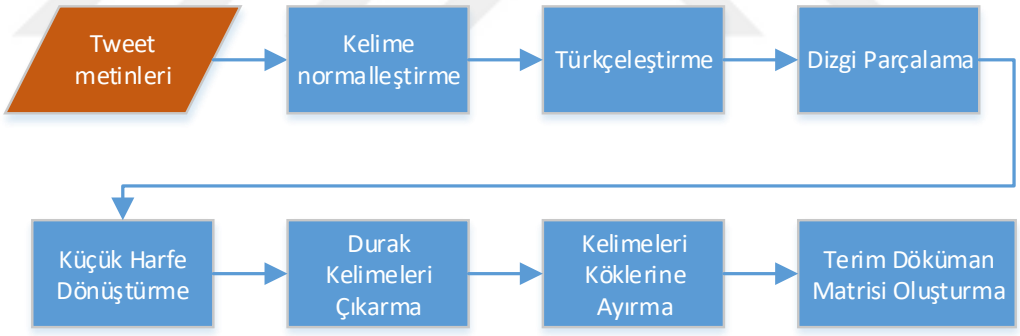
Tweetler genellikle günlük konuşma dilinde yazıldığı için gürültülü bir yapıya sahiptirler. Kelimelerin yanlış yazılması, uzatılması ya da kısaltılması, sosyal medya jargonunun kullanılması ya da Twitter'a ait özel ifadelerin kullanılması bunun başlıca nedenleridir. Bu nedenle tweetler üzerinde metin sınıflandırma işlemi öncesinde öznitelik çıkarımı yapılabilmesi için öncelikle metinlerin ön işlemeye tabii tutulup temizlenmesi gerekmektedir. Bu sayede sınıflandırma algoritmasının ayrım yapabileceği özelliklere dönüştürülerek başarılı bir sınıflandırıcı elde edilebilir. Tweet metinlerine uygulanan ön

işlemler Şekil 14'de gösterildiği gibi sırası ile kelimeleri normal hale getirme, Türkçeleştirme, parçalama, küçük haf yapma, durak kelimelerin çıkarılması, kelimelerin köklerine ayrılması ve TF-IDF vektörünün oluşturulmasıdır.

- Kelime normalleştirme: kelimelerin günlük konuşma dili şeklinde yazılması ya da karakter sınırları yüzünden kısaltılması gibi nedenlerden dolayı bazı kelimeler olduğundan farklı şekilde yazılabilmektedir. "Çok" kelimesi yerine "çoook", "seviyorum" kelimesi yerine "svyrm" şeklinde yazılan kelimelerin normalleştirilmesi için python program dili kullanılmış ve bu şekilde tekrarlayan ya da ünsüz harflerin tekrarlandığı kelimeler belirlenip düzeltilmiştir.
- Türkçeleştirme: Türkçe karakter kullanılmamasından dolayı aynı kelimenin farklı yazımları bulunabilmektedir; çıkıyor yerine cikiyor yazılması gibi. Bu şekildeki kelimelerin karakter düzeltme işlemi python programlama dili ile yapılmış ve tüm veri seti Türkçeleştirilmiştir.
- Dizgi Parçalama (Tokenize): Dizgi parçalama işlemi tweetleri kelime parçalarına ayırmakta kullanılır. Bu sayede veriyi parçalara ayırıp anlamlı bilgiler elde etmek hedeflenir. Bu işlem ile tweette bulunan her kelime ayrı nesnelere dönüştürülmüş olur. Bu sırada sayılar ve noktalama işaretleri, URL, hastag vb. kelime olamayan karakterler de kaldırılır. Örneğin "Bugün hava çok güzel" cümlesi parçalama işleminden sonra "bugün", "hava", "çok", "güzel" şeklinde 4 farklı nesne haline gelir.
- Küçük harf yapma: bu işlem ile metinlerdeki tüm harfler küçük harf haline getirilir. Örneğin "Bugün hava çok GÜZEL" cümlesi işlem sonunda "bugün hava çok güzel" şekline dönüşür.
- Durak kelimelerin çıkartılması: her dilde cümle bütününde anlamı bulan ama tek başına anlamı olmayan kelimeler vardır. Ama, ve, ise, için gibi kelimeler durak kelimelerdir. Bunlar öznitelik olarak fayda sağlamadıkları için ön işleme safhasında temizlenirler. Çalışmada literatürde kullanılan durak kelimeler incelenmiş ve buna göre bir liste oluşturulmuştur (Ek 2).

- Kelimelerin köklerine ayrılması (Stem): Öznitelik boyutunun düşürülmesi için fayda sağlayan kelimeleri köküne indirgeme işlemidir. Bu sayede kelimeler yalın hale getirilmiştir.
- Terim doküman matrisi oluşturulması: bir terimin doküman içerisinde ne kadar önemli olduğunu gösteren ağırlıklarının belirlenmesi işlemidir. Öznitelik ağırlıkları çeşitli tekniklerle yapılabilmektedir. Bunlardan bazıları TF (Terim Frekansı), Boolean, ve TF-IDF (Terim Frekansı-Ters Doküman Frekansı)'dir. Farklı ağırlıklandırma modelleri kullanan tekniklerin ortak amacı özniteliklerin ağırlıklarının gösterildiği terim doküman vektörlerinin elde edilmesidir. Çalışmada TF-IDF adı verilen terimlerin kaç kere ve kaç farklı dokümanda geçtiğini beraber araştıran teknik kullanılmıştır.

Şekil 14: Metin ön işleme süreci



Bu tez çalışması için toplanan tüm tweetlerdeki metin bilgileri birleştirilerek tek bir dosya haline getirilerek ön işlemeye hazır hale getirilmiştir. Fakat çalışmada duygu analizi yapılacak olması nedeniyle, kişilerin duygularını ifade etmek için sıklıkla kullandıkları emoji adı verilen ifadelerin noktalama işaretleri ile beraber silinmemesi önem arz etmektedir. Bu nedenle ön işleme sürecine başlamadan önce bu ifadeler tespit edilmiş ve pozitif ifade ve negatif ifade olarak etiketlenmiştir.

Bahsedilen ön işleme yapılabilmesi amacıyla "Rapidminer" ve "R" programları kullanılmıştır. Rapidminer kolay anlaşılır ve uygulanabilir bir ara yüze sahip olması nedeniyle, R ise açık kaynak kodlu bir programlama dili olması sebebiyle tercih

edilmiştir. İki program da gerek büyük hacimli veri setlerinde gösterdikleri performans gerek makine öğrenmesi ve veri madenciliği alanında etkin sonuçlar sağlaması nedeniyle ön plana çıkmaktadır. Yukarıda bahsedilen işlemler için Rapidminer programında "Process Document from Data" operatörüne dizgi parçalama, küçük harf yapma, durak kelime çıkarma ve köklere ayırma alt operatörlerinin eklenmesi ile bir ön işleme süreci hazırlanmıştır. R programında ise "tm" kütüphanesi kullanılmış ve dizgi parçalama, küçük harf yapma, durak kelime çıkarma ve köklere ayırma işlemlerin komutları eklenmiştir. Bu işlemler sonunda her iki program ile de TF-IDF matrisi elde edilmiştir. Metin ön işleme süreçlerinin Rapidminer programında uygulanışı ve R program kodları ve terim doküman matrisi örneği Ek 3'de paylaşılmıştır.

5.3. Duygu Analizi

Kişiler konuşurken ya da yazarken genel olarak iki şeyi amaçlarlar; gerçekleri belirtmek ya da görüşlerini bildirmek. Var olan herkesçe bilinen durumlar üzerine konuşulması ya da bu konular hakkında yazılması gerçekleri belirtmek amaçlıdır. Görüş bildirme amaçlı yapılan konuşmalar ya da yazılan iletiler ise kişiden kişiye değişen, kişinin o anki ruh halini, ilgili konu hakkındaki düşüncesini yansıtmaktadırlar. Duygu analizi görüş bildirilen bu iletilerin pozitif ya da negatif duygular ile ya da mutlu, üzgün, umutsuz vb. ruh halleriyle yazıldığının otomatik tespit edilmesini amaçlamaktadır. Bu çalışmada Twitter'dan ilgili kanal ve o kanalda 8 aylık periyodda yayınlanan programlar hakkında yazılan iletiler toplanmıştır. Bu iletilerden izleyici görüşlerinin belirlenebilmesi amacıyla duygu analizi yapılması hedeflenmiştir. Böylece izleyicilerin ilgili kanalı ve programı izlerken olumlu mu olumsuz mu düşündükleri ortaya çıkarılacak ve bu bilgiler reyting tahminlerinde kullanılarak izleyici görüşleriyle desteklenmiş bir tahmin modeli kurulacaktır.

R ve python gibi programlama dillerinde; yabancı diller için geliştirilmiş bir çok duygu analizi kütüphanesi bulunmaktadır. Türkçe için geliştirilmiş böyle bir kütüphane ya da yazılım var olmamasına karşın bu alanda çok sayıda çalışma yapılmakta, çeşitli gruplarca yüksek performanslı doğal dil işleme ve duygu analizi modülleri geliştirilmeye

çalışılmaktadır. Bu nedenle literatürdeki duygu analizi çalışmalarının tamamına yakını mühendislik temelli araştırmacılar tarafından yapılmaktadır. Bu çalışmada ise sosyal bilimcilerin de duygu analizi tekniğini kullanabilmeleri için daha kolay kullanılabilen platformlarda, göreceli olarak uygulaması daha kolay olan bir duygu analizi algoritması sunulmaktadır. Öncelikle geliştirilen yarı denetimli sözlük temelli duygu analiz algoritması ile 1.200.000 tweetin duygusu belirlenmeye çalışılmıştır. Sonrasında eğitim seti olarak sözlük temelli yöntem ile belirlenen verilerin kullanıldığı bir makine öğrenmesi temelli duygu analizi modeli kurulmuş ve tahmin performansı incelenmiştir.

5.3.1. Önerilen Duygu Analiz Algoritması

Sözlük temelli duygu analizi algoritmalarında kritik konulardan biri sözlük oluşturma aşamasıdır. Her veri seti kendine has jargonlara sahip olabilmektedir görüşünden yola çıkarak bu çalışmada yarı denetimli bir teknikle sözlük oluşturulmuştur. Bu amaçla 1000 adet tweet pozitif olarak 1000 adet tweet de negatif olarak etiketlenmiştir. Sonrasında her iki veri seti için de terim frekansları belirlenmiş ve;

- pozitif etiketli tweetlerde en sık geçen 1000 kelime
- negatif etiketli tweetlerde en sık geçen 1000 kelime

şeklinde iki farklı kelime seti elde edilmiştir. Bu veri setleri içerisindeki anlamlı kelimeler seçilerek "Pozitif Duygu Sözlüğü" ve "Negatif Duygu Sözlüğü" oluşturulmuştur. Ardından veri setindeki tüm iletilerin pozitif duygu sözlüğünden ve negatif duygu sözlüğünden kaçar kelime içerdikleri belirlenmiş ve sayıca yüksek olan duyguya etiketlenmiştir.

Duygu sözlüklerinin ve veri setinin küçük örnekleri üzerinden algoritmanın işleyişi ve adımları Tablo 7, Tablo 8, Tablo 9 ve Tablo 10'da gösterilmektedir.

Tablo 7: Duygu sözlükleri örneği

Pozitif duygu sözlüğü	güzel, mükemmel, iyi ….
Negatif duygu sözlüğü	kötü, çirkin …

Tablo 8: Pozitif duygu sözlüğü kontrolü ve skor belirleme (adım 1)

	güzel	mükemmel	iyi	…	Skor
Bu program çok iyi, sunucu da güzel	1	0	1		2
Ne kadar çirkin dizi	0	0	0		0
Bu akşam benim dizim var	0	0	0		0

Tablo 9: Negatif duygu sözlüğü kontrolü ve skor belirleme (adım 2)

	kötü	çirkin	…	Skor
Bu program çok iyi, sunucu da güzel	0	0		0
Ne kadar çirkin dizi	0	1		1
Bu akşam benim dizim var	0	0		0

Tablo 10: Duygu durumu belirleme için skor karşılaştırma (adım 3)

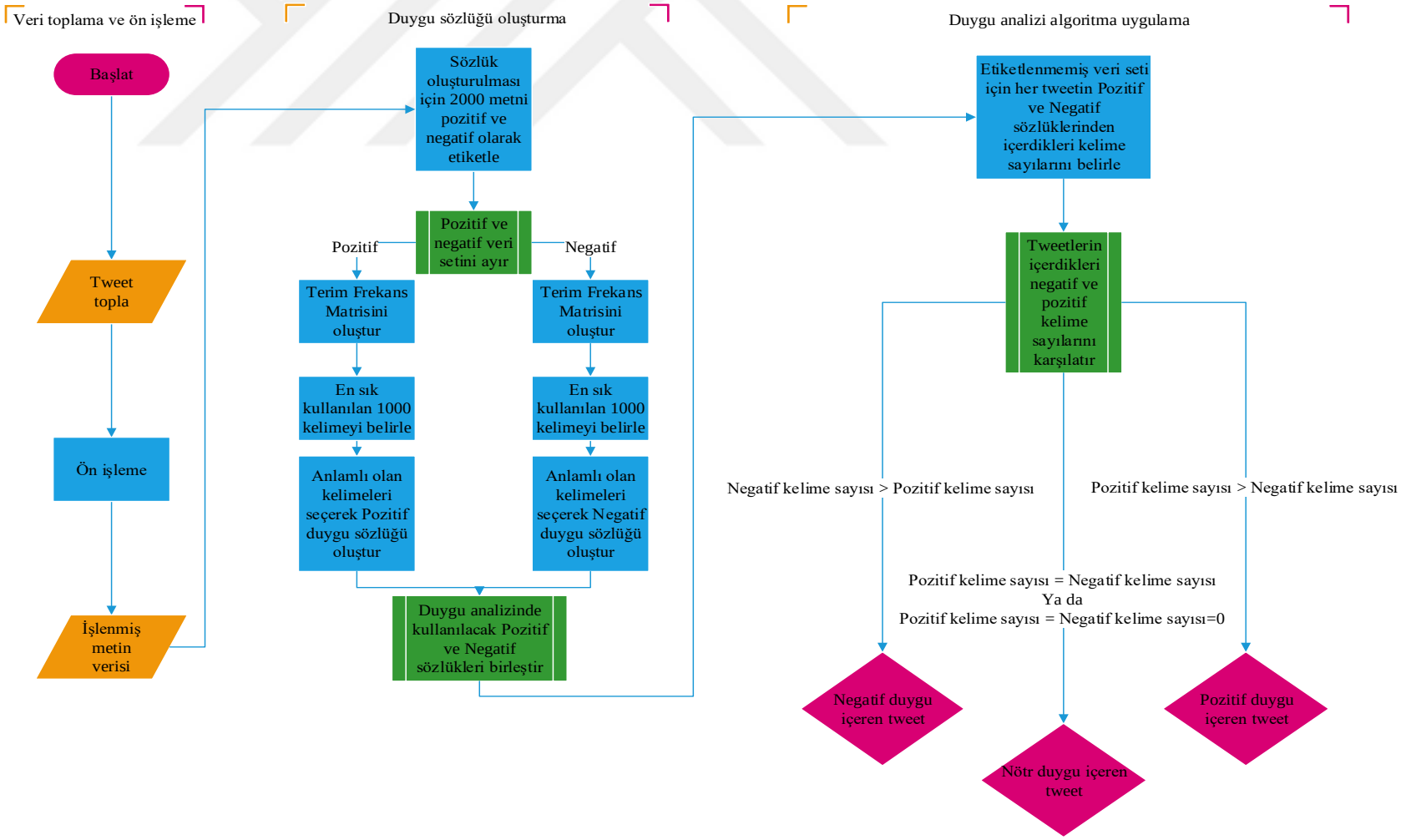
	Poz. sözlük skoru	Neg. sözlük skoru	Sonuç
Bu program çok iyi, sunucu da güzel	2	0	Pozitif
Ne kadar çirkin dizi	0	1	Negatif
Bu akşam benim dizim var	0	0	Nötr

Sözlük temelli yaklaşım ile yapılan bu duygu analizi çalışmasının yarı denetimli olarak ifade edilmesinin nedeni sözlüklerin oluşturulurken etiketlenmiş veri setlerinden yararlanılmış olmasıdır. 1.200.000 tweetlik veri setinin bu yöntem ile duygu durumunun belirlenmesi işleminin ardından bu makine öğrenmesi temelli yöntemlerle algoritmanın performansı test edilmiştir. Bu amaçla veri seti çapraz değerlendirme yapılabilmesi için 9 parçaya ayrılmış ve her bir parça da eğitim ve test olmak üzere ikiye ayrılmıştır. Daha önceki aşamalarda bahsedilen metin ön işleme süreçlerinin uygulanması ile veri setlerinin analize hazır hale getirilmiş ve tweetlerin duygularının tahmin edilmesi amacıyla karar ağacı algoritması kullanılmıştır. Böylece sözlük temelli duygu analizi algoritmasının tahmin performansları şu şekilde belirlenmiştir;

Veri seti	Doğruluk
1	%69
2	%66,8
3	%67,4
4	%69,18
5	%67,73
6	%67,11
7	%70,18
8	%68,97
9	%67,22
Ortalama	%68.12

Literatürde yer alan yoğun programlama bilgisi temelli duygu analizi çalışmaları daha iyi performanslar ile sonuçlar üretmektedir ve birçok mühendislik temelli araştırmacı bu performansları arttırmaya yönelik algoritmalar üzerine çalışmaktadır. Fakat ilgili çalışmalar uygulama zorluğu nedeniyle her alandan araştırmacıların kullanabileceği yapıya sahip değillerdir. Bu da metinlerde gizli kalmış duygularının açığa çıkarılması ile yaratılacak değerlerden yoksun kalınmasına neden olmaktadır. Şekil 15'de detaylıca görselleştirilen, tez çalışması kapsamında geliştirilen bu model ise uygulama kolaylığı ön planda olan performans kalitesi yüksek bir duygu analizi algoritması önermektedir. Modelde sözlük oluşturma aşamasında R programlama dili, sözlük skoru belirleme ve karşılaştırma aşamalarında ise Excel ve VBA kullanılmıştır. Önceki bölümlerde de açıklandığı gibi R programlama dili günümüzde veri analitiği çalışmalarında sıklıkla kullanılan uygulama kolaylığı sunan bir yazılımdır. Excel ise Microsoft Office programlarından biri olarak milyonlarca bilgisayarda yer alan elektronik tablolama programıdır. Bu da birçok araştırmacı tarafından daha kolay kullanım imkânı sunmaktadır. Kullanılan tüm kodlar ve fonksiyonlar Ek 4'te yer almaktadır.

Şekil 15: Duygu analizi algoritması



5.4. Numerik Veri Ön İşleme

Çalışmanın bu aşamasında, Twitterden çekilen ham verilerin düzenlenmesi, duygu analizi sonuçlarının mevcut veriler ile birleştirilmesi, dakikalık reytinglere ile dakikalık tweetlerin bir araya getirilmesi gibi veri setinin veri madenciliğine hazır hale getirilmesi için yapılan işlemleri açıklanacaktır. Elde edilen tweet veri setinin saklanmasında ilişkisel veri tabanı yönetim sistemi olan SQL yazılımı kullanılmıştır. Uygulama esnasında kullanım kolaylığı olması adın Excel programına ilgili veriler çağırılarak Excel üzerinden de çeşitli sorgular yapılmıştır. Her tweete mevcut olan ait kullanıcı adı, tarih, metin, retweet sayısı, favori sayısı bilgilerine, TV programı adı ve duygu durumu etiketleri eklendi. İlgili tweetler hangi TV programı adı ile sorgulanıp toplandıysa program adı sütununa bilgisi girildi. Öte yandan duygu analizi sonucu belirlenen duygu sınıfı duygu durum sütununda belirtilmiş. Böylece her tweet için "x TV programı hakkında, x kişisi tarafından, x zamanda, x metin içeriğine sahip, x kere retweet edilmiş, x kere favori edilmiş ve x duygusunu barındırmaktadır" bilgisi derlenmiştir.

Öte yandan kanaldan alınan reyting veri setinde ve bunlardan elde edilen yayın akışlarında program adlarının yanında tekrar ya da özet gibi ek bilgiler yazmaktaydı. Seyircilerin özet izliyorlarken tweet atması durumunda özet yazmaya ihtimalleri çok düşük olduğu için yayın akışındaki program adları sadeleştirilerek sorgularda kolaylık sağlaması hedeflenmiştir.

Her ayın dakikalık yayın akışı, ilgili ay için kanal adı ile yapılan tweet sorgu sonuçları ve ilgili ay yayın akışında olan programlar için program adı ile yapılan tweet sorgu sonuçları her birinin farklı çalışma sayfalarında yer aldığı Excel dosyalarında toplanmıştır. Böylece "dakikalık reyting", "tweet-kanal adı", "tweet-program adı" şeklinde 3 çalışma sayfası barındıran; Kasım.xlsx, Aralık.xlsx şeklinde 8 farklı dosya elde edilmiştir. Dakikalık reyting çalışma sayfalarında her dakika için;

- İlgili dakikada ilgili program hakkında atılan tweet sayısı
- Atılan tweetlerin kaçı nötr duygulu
- Atılan tweetlerin kaçı pozitif duygulu
- Atılan tweetlerin kaçı negatif duygulu
- Atılan nötr duygulu tweetlerin retweet edilme sayısı
- Atılan pozitif duygulu tweetlerin retweet edilme sayısı
- Atılan negatif duygulu tweetlerin retweet edilme sayısı
- Atılan nötr duygulu tweetlerin favori edilme sayısı
- Atılan pozitif duygulu tweetlerin favori edilme sayısı
- Atılan negatif duygulu tweetlerin favori edilme sayısı
- İlgili dakikada ilgili kanal hakkında atılan tweet sayısı
- Atılan tweetlerin kaçı nötr duygulu
- Atılan tweetlerin kaçı pozitif duygulu
- Atılan tweetlerin kaçı negatif duygulu
- Atılan nötr duygulu tweetlerin retweet edilme sayısı
- Atılan pozitif duygulu tweetlerin retweet edilme sayısı
- Atılan negatif duygulu tweetlerin retweet edilme sayısı
- Atılan nötr duygulu tweetlerin favori edilme sayısı
- Atılan pozitif duygulu tweetlerin favori edilme sayısı
- Atılan negatif duygulu tweetlerin favori edilme sayısı
- dakikanın 5 dakika öncesi ve 5 dakika sonrasında ilgili program hakkında atılan toplam tweet sayısı
- aralıkta atılan tweetlerin kaçı negatif duygulu
- aralıkta atılan tweetlerin kaçı pozitif duygulu
- aralıkta atılan tweetlerin kaçı nötr duygulu
- dakikanın 5 dakika öncesi ve 5 dakika sonrasında ilgili kanal hakkında atılan toplam tweet sayısı
- aralıkta atılan tweetlerin kaçı negatif duygulu

- aralıkta atılan tweetlerin kaçı pozitif duygulu
- aralıkta atılan tweetlerin kaçı nötr duygulu
- dakikanın 10 dakika öncesi ve 10 dakika sonrasında ilgili program hakkında atılan toplam tweet sayısı
- aralıkta atılan tweetlerin kaçı negatif duygulu
- aralıkta atılan tweetlerin kaçı pozitif duygulu
- aralıkta atılan tweetlerin kaçı nötr duygulu
- dakikanın 10 dakika öncesi ve 10 dakika sonrasında ilgili kanal hakkında atılan toplam tweet sayısı
- aralıkta atılan tweetlerin kaçı negatif duygulu
- aralıkta atılan tweetlerin kaçı pozitif duygulu
- aralıkta atılan tweetlerin kaçı nötr duygulu

şeklinde 36 sorgu yapılmıştır. Sorgular için aynı dosyada bulunan diğer çalışma sayfaları kullanılarak "eğersay" ve "eğertopla" komutları ile veriler elde edilmiştir. Bu sayede elde edilen aylık dosyalara ait görselleri Ek 5'te paylaşılmıştır.

Düzenlenen aylık dosyaların birleştirilmesi ile 8 aylık döneme ait reyting, atılan tweet, tweetlerin duygu durumları bilgilerinin ve yukarıda bahsedilen sorgu sonuçlarının yer aldığı "genel veri seti" adında tek bir dosya oluşturulmuştur. Bu dosyada 343.517 satır, 46 satır bulunmaktadır. Numerik hale çevrilmiş bu veri setinin analize uygun hale gelebilmesi için bazı temizleme işlemleri yapılmıştır. Toplamda 34.347 satırlık veri ilgili program için ilgili dakikada tweet çekilemediği için silinmiştir. Bunun yanı sıra uygulamada ihtiyaç duyulmamasından dolayı program adı, tarih ve saat sütunları da silinmiştir. Böylece kullanıma hazır hale gelen "dakikalık reyting veri seti" 309.170 satır, 43 sütun şeklinde sadeleştirilmiştir.

Bir önceki adım ile eş zamanlı olarak, "genel veri seti" dosyasından Excel'in "düşeyara", "eğer", "eğersay" ve "eğertopla" komutları kullanılarak 24 farklı sorgu yapılarak aşağıdaki bilgileri içeren 3921 satır ve 32 sütunluk "program reyting veri seti" oluşturulmuştur (Ek 6).

- 8 aylık periyodda yayınlanmış program adları
- Yayın tarihi
- Yayın başlangıç saati
- Yayın bitiş saati
- Yayının reyting değerleri
- Yayın süresince program adı ile atılan toplam tweet sayısı
- Yayın süresince program adı ile atılan toplam tweetlerin kaçı nötr duygulu
- Yayın süresince program adı ile atılan toplam tweetlerin kaçı pozitif duygulu
- Yayın süresince program adı ile atılan toplam tweetlerin kaçı negatif duygulu
- Program adı ile atılan nötr duygulu tweetlerin retweet edilme sayısı
- Program adı ile atılan pozitif duygulu tweetlerin retweet edilme sayısı
- Program adı ile atılan negatif duygulu tweetlerin retweet edilme sayısı
- Program adı ile atılan nötr duygulu tweetlerin favori edilme sayısı
- Program adı ile atılan pozitif duygulu tweetlerin favori edilme sayısı
- Program adı ile atılan negatif duygulu tweetlerin favori edilme sayısı
- Yayın süresince kanal adı ile atılan toplam tweet sayısı
- Yayın süresince kanal adı ile atılan toplam tweetlerin kaçı nötr duygulu
- Yayın süresince kanal adı ile atılan toplam tweetlerin kaçı pozitif duygulu
- Yayın süresince kanal adı ile atılan toplam tweetlerin kaçı negatif duygulu
- Kanal adı ile atılan negatif duygulu tweetlerin retweet edilme sayısı
- Kanal adı ile atılan pozitif duygulu tweetlerin retweet edilme sayısı
- Kanal adı ile atılan nötr duygulu tweetlerin retweet edilme sayısı
- Kanal adı ile atılan negatif duygulu tweetlerin favori edilme sayısı
- Kanal adı ile atılan pozitif duygulu tweetlerin favori edilme sayısı
- Kanal adı ile atılan nötr duygulu tweetlerin favori edilme sayısı
- Program başlangıç saatinden 2 saat önce ve bitişinden 2 saat sonra aralığında program adı ile atılan toplam tweet sayısı

Bunlara ek olarak tahmin modelinde ilk altısı girdi ve sonuncusu çıktı olarak kullanılmak üzere, “program reyting veri seti”ne aşağıdaki bilgileri içeren sütunlar eklenmiştir;

- Hitap ettiği kitle
- Yayın aralığı
- Program türü
- Geçen sezon var mıydı?
- Programdaki ünlü seviyesi
- Yayın günü
- Gelecek bölüm reyting sınıfı

Güncellenen program reyting veri setinin tahmin modellerinde kullanıma hazır olması için eksik ve işe yaramayacak olan bilgiler temizlenmiştir. Bu amaçla yayın aralığında tweet çekilememiş programları içeren 217 satır, çocukların tweet atabilecek grupta yer almamalarından dolayı çizgi film türündeki programları içeren 433 satır, sonraki bölüm reyting tahmin çalışmasında kullanılması mümkün olmadığı için sinema filmlerini içeren 121 satır ve ilgili programların final bölümü olan 75 satır silinmiştir. Böylece tekrar düzenlenen program reyting veri setinde 3075 satır veri kalmıştır.

Tahmin modellerinde çıktı olarak kullanılmak üzere düzenlenen, gelecek bölüm reyting sınıfı, gerçek reyting verilerinde yer alan 3 farklı sosyo ekonomik statüye göre belirlenmiştir. Böylece tahmin modelleri her üç çıktı değeri için de ayrı ayrı uygulanacaktır.

Böylece veri madenciliği çalışmalarının yaklaşık %70-75’ini oluşturan veri setinin hazırlanması evresi sonlandırılmış ve çeşitli yöntemlerin uygulanabileceği temizlenmiş veri setleri oluşturulmuştur.

5.5. Analiz

5.5.1. Kümeleme

Kümeleme analizi basitçe veri setindeki birbirine benzer nesnelerin gruplandırılması olarak ifade edilebilir. Kümeleme analizi denetimsiz öğrenme tekniği ile veri setindeki gizli örüntülerin araştırılmadı olarak da açıklanabilir (Akpınar, 2014).

309.170 satırlık dakikalık reyting veri seti içerisinde dakikalık olarak hesaplanmış gerçek reyting değerlerini ve her dakika için atılan tweet sayısı, tweetlerin duyguları, duygularına göre retweet ve favori edilme sayılarını içermektedir. Gerçek reyting değerlerinin ayrıştırıldığı veri setine kümeleme analizi yapılmıştır. İdeal küme sayısını belirleyebilmek için çeşitli k değerleri denenmiş ve Tablo 11'de görülebileceği gibi en iyi Davien-Bouldin endeks değerini veren k değeri analizde kullanılmak üzere seçilmiştir. Kümeleme analizi ile elde edilen küme değerlerini kullanılarak; dakikalık tweet sayıları ve bu tweetlerin duygularının oluşturduğu gruplar ile gerçek reyting sınıfları arasında ilişki olup olmadığı araştırılmıştır. Gelecek bölüm reytinglerinin tahminlenmesi için kurulacak modelde sonuç/ çıktı değeri olan reyting sınıfı kanal yetkilileri tarafından verilen aralıklara göre düzenlenmiştir. Kümeleme analizi sonuçları ile gerçek reyting sınıflarının birbirinden bağımsız değişkenler olup olmadığı ki-kare bağımsızlık testi uygulanarak incelenmiştir.

Tablo 11: Kümeleme sonuçları

Küme sayısı	Davies-Bouldin Endeksi
3	-0,693
4	-0,662
5	-0,710
6	-0,586
7	-0,593
8	-0,531

En iyi Davies-Bouildin endeks değerini veren küme sayısı 5 olduğu için k değerinin 5 olduğu kümeleme analiz sonuçları incelenecektir. Buna göre 1. küme 292157, 2. küme 9817, 3. küme 214, 4. küme 5694 ve 5. küme 1287 veri içermektedir. Kümeleme analizinin sonuçları incelendiğinde her kümedeki nesne sayısının yani küme

büyüklüklerinin birbirinden farklı olduğu görülmektedir. Ancak bu durum küme analizinin sonuçlarının doğruluğu ile ilgili bir durum değildir. Kümeleme sonuçları ile tweet veri setine göre birbirine benzeyen zamanlar gruplanmıştır. Gerçek reyting veri setinin uzman görüşü doğrultusunda kategorize edilmesi de sayısal anlamda birbirine yakın değerlerin gruplanmasını ifade etmektedir. Bu sayede aynı dakikalar iki farklı yöntem ile gruplandırılmıştır ve bu yöntemler karşılaştırılmak üzere aşağıdaki hipotez test edilmiştir.

H_0: Küme sonuçları ile gerçek reyting sınıfları birbirinden bağımsız değişkenlerdir ve iki değişken arasından bir ilişki yoktur.

H_1: Küme sonuçları ile gerçek reyting sınıfları arasından bir ilişki vardır.

Hipotezin test edilmesi için ki-kare bağımsızlık testi uygulanmış ve $p<0.05$ anlamlılık düzeyi seçilmiştir. Elde edilen test sonuçlarından p değeri 0.00 bulunmuş bu da küme sonuçları ile gerçek reyting sınıfları arasında bir ilişki olduğunu bunların birbirinden bağımsız değişkenler olmadığını göstermiştir.

5.5.2. Sınıflandırma

Çalışmanın bu aşamasında veri madenciliği ve işletme analitiğinde önem arz eden, veriden anlamlı sonuçlar çıkartarak değer yaratma, işletmeler için yeni stratejiler geliştirecek, kararlar almasını sağlayacak bilgiler üretme amacı üzerinde durulmuştur. Büyük veri madenciliğini istatistikten ayıran en önemli özellik veriden yola çıkarak araştırma sorularının oluşturulmasıdır. Bu amaçla, çalışmada sosyal medya verilerinin işletmeler için öneminin, kullanılabilirliğinin ve anlamlandırılabilir olmasının ortaya konulması fikrinden yola çıkılmıştır. Geliştirilen tahmin modeli ile literatürde kullanılan girdilere ek olarak tweet verilerinin nitelik ve nicelik değerleri de kullanılmıştır. Buna göre oluşturulan araştırma sorusu şu şekildedir;

“Bir işletmenin (TV kanalı), ürünlerine (TV programları) has özellikler (program türü, yayın saati vb.) ve sosyal medyada (Twitter) işletme ve ürünleri hakkındaki konuşmaların

sayısı ve bu konuşmaların içerdikleri duygu durumları ile gelecek dönemdeki satış değeri tahmin edilebilir mi?"

Denetimli öğrenme tekniklerinden karar ağacı, destek vektör makineleri ve yapay sinir ağları algoritmaları kullanılarak araştırma sorusuna yönelik sınıflandırma modelleri kurulmuştur. Kurulan modellerde Tablo 12'de detaylıca açıklanan değişkenler kullanılmıştır.

Tablo 12:Tahmin modelinde kullanılan değişkenler

Türü	Kodu	Değişken adı	Değerler
Girdi Değişkeni	x1	Programın hitap ettiği kitle	1-genel izleyici kitlesi 2-13+kadın 3-13+
	x2	Programın yayınlandığı aralık	1- 20:00-22:59 2- 07:00-19:59 3- 23:00-06:59
	x3	Programın türü	21-Yorum Programı 22-Sohbet Programı 31-Belgesel Programlar 32-Bilgi-Kültür Yarışmaları 43-Bilgi-Beceri Programları 51- Reality Show 76-Beceri ve Direnç Yarışmaları
	x4	Program geçen sezon yayında mıydı?	0- Hayır 1- Evet
	x5	Programdaki ünlü seviyesi	1- Çok 2- Orta 3- Az
	x6	Program hangi gün yayınlanıyor	1- Pazartesi 2- Salı 3- Çarşamba 4- Perşembe 5- Cuma

			6- Cumartesi 7- Pazar
	x7	Program süresince program adı ile atılan toplam tweet sayısı	Max- 31933 Min- 0
	x8	Program sürecince program adı ile atılan Nötr duygulu tweetlerin sayısı	Max- 20255 Min- 0
	x9	Program sürecince program adı ile atılan Pozitif duygulu tweetlerin sayısı	Max- 6547 Min- 0
	x10	Program sürecince program adı ile atılan Negatif duygulu tweetlerin sayısı	Max- 5879 Min- 0
	x11	Program süresince kanal adı ile atılan toplam tweet sayısı	Max- 1568 Min- 0
	x12	Program sürecince kanal adı ile atılan Nötr duygulu tweetlerin sayısı	Max- 891 Min- 0
	x13	Program sürecince kanal adı ile atılan Pozitif duygulu tweetlerin sayısı	Max- 726 Min- 0
	x14	Program sürecince kanal adı ile atılan Negatif duygulu tweetlerin sayısı	Max- 726 Min- 0
	x15	Program süresince program adı ile atılan Nötr duygulu tweetlerin retweet edilme sayısı	Max- 20842 Min- 0
	x16	Program süresince program adı ile atılan Pozitif duygulu tweetlerin retweet edilme sayısı	Max- 11633 Min- 0
	x17	Program süresince program adı ile atılan Negatif duygulu tweetlerin retweet edilme sayısı	Max- 5167 Min- 0
	x18	Program süresince program adı ile atılan Nötr duygulu tweetlerin favori edilme sayısı	Max- 112255 Min- 0
	x19	Program süresince program adı ile atılan Pozitif duygulu tweetlerin favori edilme sayısı	Max- 34578 Min- 0
	x20	Program süresince program adı ile atılan Negatif duygulu tweetlerin favori edilme sayısı	Max- 35541 Min- 0
	x21	Program süresince kanal adı ile atılan Nötr duygulu tweetlerin retweet edilme sayısı	Max- 6430 Min- 0
	x22	Program süresince kanal adı ile atılan Pozitif duygulu tweetlerin retweet edilme sayısı	Max- 9201 Min- 0
	x23	Program süresince kanal adı ile atılan Negatif duygulu tweetlerin retweet edilme sayısı	Max- 1582 Min- 0

	x24	Program süresince kanal adı ile atılan Nötr duygulu tweetlerin favori edilme sayısı	Max- 6594 Min- 0
	x25	Program süresince kanal adı ile atılan Pozitif duygulu tweetlerin favori edilme sayısı	Max- 25657 Min- 0
	x26	Program süresince kanal adı ile atılan Negatif duygulu tweetlerin favori edilme sayısı	Max- 1452 Min- 0
	x27	Program başlamadan önce 2 saat ve bittikten sonra 2 saat aralığında program hakkında atılan toplam tweet sayısı	Max- 34390 Min- 0
Çıktı Değişkeni	y	Gelecek bölüm reyting sınıfı	0.000;0.2 -> a 0.201;0.8 -> b 0.801;2 -> c 2.001;4 -> d 4.001;11 -> e

Sınıflandırma modelleri, veri setinin eğitim ve test olarak ikiye ayrılması ile eğitim setinden kullanılan algoritmaya göre veri çeşitli yöntemlerle öğrenerek test setinin sınıf değerinin tahmin edilmesini amaçlar. Model tahmin performansını değerlendirebilmek için; gerçek sınıf değerleri ile tahmin edilen sınıf değerlerinin yer aldığı, sınıf değerleri bilinen verilerin hangi tahmin sonucu sınıfa atandığını gösteren karşılaştırma matrisi (Tablo 13) düzenlenir ve aşağıdaki listelenen popüler performans değerlendirme amaçlı kullanılan değerler hesaplanır.

Tablo 13: İkili karşılaştırma matrisi

	Gerçek		
Tahmin		+	-
	+	**TP** Doğru Pozitif	**FP** Yanlış Pozitif
	-	**FN** Yanlış Negatif	**TN** Doğru Negatif

- Doğruluk - Hata Oranı; Modelin performansını ölçmede kolay ve anlaşılır olması nedeniyle doğruluk ve hata oranlarını hesaplamak sıklıkla tercih edilen bir yöntemdir. Doğruluk oranı doğru bilinen pozitif veya negatif değerlerin tüm

veriye oranıdır. Hata oranı ise gerçekte pozitif olan tüm değerlerin tüm veriye oranıdır.

Doğruluk Oranı (Accuracy) = (TP+TN) / (TP+FP+TN+FN)
Hata Oranı = (FP+FN) / (TP+FP+TN+FN)

- Kesinlik (Precision); doğru bilinen pozitif verilerin pozitif olarak tahminlenen tüm verilere oranıdır.

  Kesinlik = TP/(TP+FP)

- Duyarlılık (Sensitivity) ; doğru pozitiflerin tüm gerçek pozitiflere oranıdır.

  Duyarlılık = TP/(TP+FN)

- Özgüllük (Specificity); Doğru bilinen negatiflerin gerçek tüm negatiflere oranıdır.

  Özgüllük = TN/(TN+FP)

- F Ölçütü ; Kesinlik ve duyarlılığın harmonik ortalamasıdır.

  F Ölçütü = 2TP / (2TP + FP + FN)
  veya
  F Ölçütü = 2* [(kesinlik * duyarlılık) / (kesinlik + duyarlılık)]

Bu tez çalışmasında sınıflandırma algoritması olarak karar ağacı, destek vektör makineleri ve yapay sinir ağları kullanılmıştır. İlgili algoritmaların başarısı 10- katlamalı çapraz doğrulama yöntemi ve veri setini eğitim-test şeklinde ayrılması ile sınanmıştır. 10-katlamalı çapraz doğrulama yönteminde veri seti 10 eşit parçaya bölünür ve her seferinde

farklı bir parçası test seti olarak kullanılarak 10 defa algoritma çalıştırılmış olur ve her bir işlemde hesaplanan tahmin performansının ortalaması modelin tahmin başarısını gösterir. Veri setini eğitim-test şeklinde ayrılmasında ise veri setinin %80'i eğitim %20'si test olmak üzere ikiye ayrılması ile tahmin başarısı değerlendirilmektedir. Karar ağacı algoritması-C4.5 yöntemi kategorik ve sürekli değişkenler ile uygulanabildiği için veri seti olduğu gibi kullanılmıştır. Fakat destek vektör makineleri ve yapay sinir ağları algoritmalarında kategorik değişkenlerin kukla değişkenlere dönüştürülmesi gerekmektedir. Veri setindeki ilk altı değişken için ilgili dönüşümler yapılarak veri seti düzenlenmiştir. Ayrıca önceki bölümlerde bahsedildiği gibi gerçek reyting ölçümleri çeşitli sosyoekonomik statülere göre yapılmakta ve bu çalışmada 3 farklı sosyoekonomik statüye göre elde edilmiş sınıf değerleri bulunmaktadır. Bu nedenle öznitelik değerleri aynı olmakla birlikte sınıf vektörünün değiştiği 3 farklı veri seti düzenlenmiştir. Sınıflandırma algoritmaları 3 veri seti için de ayrı ayrı uygulanacaktır.

5.5.2.1. Karar Ağacı

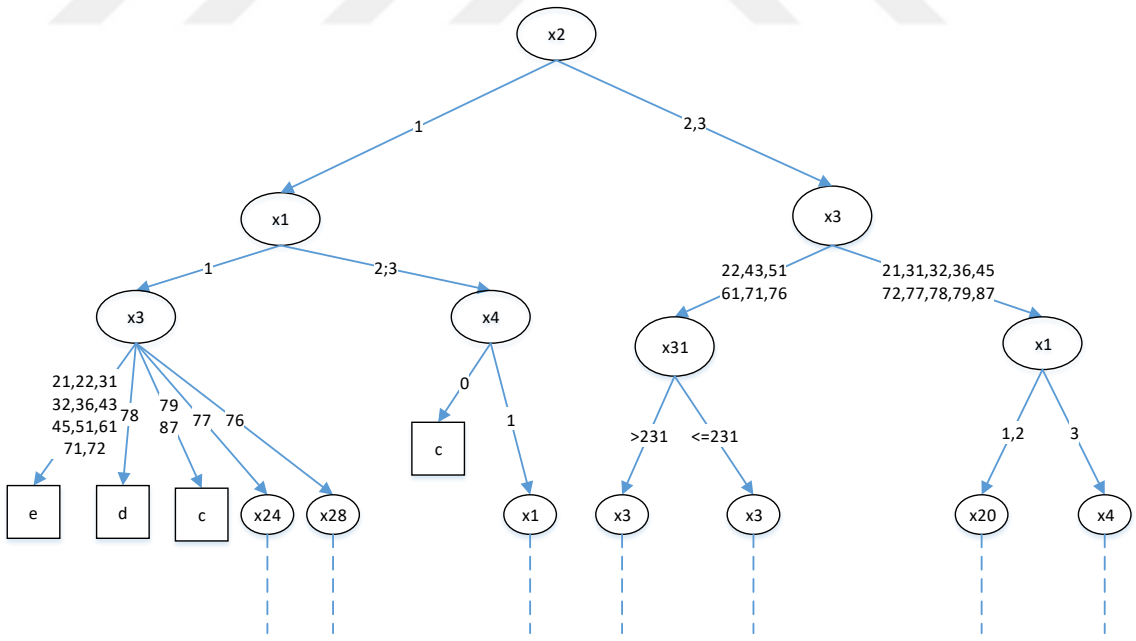
Karar ağacı algoritması bir grup karar düğümünün dallarla birbirine bağlandığı ve kök düğümden yaprak düğüme bir dizi kuralla ulaşılmasını sağlayan ağaç yapısındaki sınıflandırma algoritmasıdır. Anlaşılma ve uygulama anlamında kullanıcılara kolaylık sağlaması karar ağacı algoritmasının 1970'li yıllardan beri popüler olmasını sağlamıştır. "Eğer"…"İse" … şeklinde kurulan kurallar bütünü ile kökten yaprağa ulaşılması hedeflenen karar ağaçlarında genel anlamı ile bölme ve budama işlemleri yapılmaktadır. Önceden belirlenen bölme endeksleri vasıtasıyla en iyi bölen öznitelikler belirlenir ve bunlara göre veri seti alt kümelere ayrılır. Yinelemeli olarak devam eden bu işlem durdurma kriterine ulaşıldığında sonlanır. Budama işlemi ise karar ağacının genelleştirilebilmesi için istenmeyen alt dalların ya da düğümlerin ayıklanmasıdır.

Ross Quinlan 1970'li yılların sonunda ortaya çıkardığı ağaç temelli model, ID3'yi 80'lerde geliştirerek C4.5 modelini ardından da biraz daha yenilik ekleyerek C5.0 modelini geliştirmiştir. C5.0'ın temel özellikleri en iyi bölen özniteliğini belirlemek için entropiyi hesaplayıp bilgi kazancı değerlerini kullanması, kategorik ve sürekli

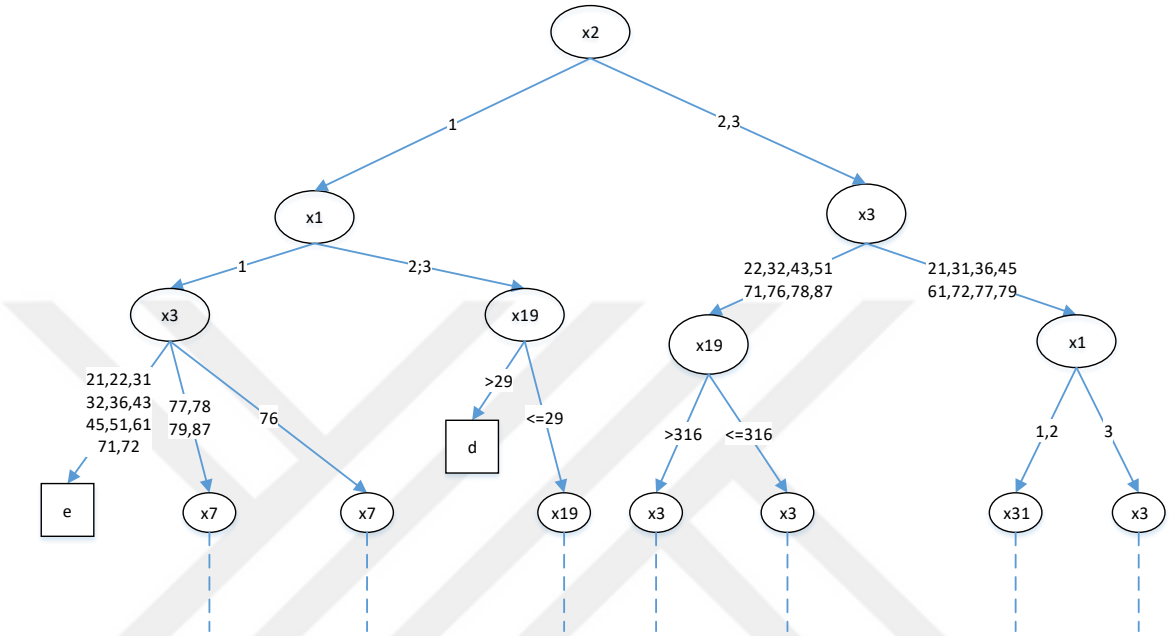
değişkenlerde kullanılabilmesi, büyük veri setlerinde hızlı uygulama şansı tanıması ve hafızada az yer kaplamasıdır. C5.0 algoritması boosting algoritması adı verilen her bir sınıflandırıcıda yapılan sınıflandırma hatasına göre örneklerin ağırlıklarının güncellendiği yöntem sayesinde doğruluğu daha yüksek bir karar ağacı oluşturulması hedeflenmektedir. Bu tez çalışmasındaki veri seti kategorik ve sürekli değişkenleri beraber barındırması ve C5.0'ın sağladığı diğer avantajlar nedeniyle karar ağacı algoritması olarak C5.0 kullanılmıştır.

27 öznitelik bir sınıf değişkenine ait 3075 farklı örneğin bulunduğu veri seti %80'i eğitim %20'si test seti olmak üzere ikiye ayrılmıştır. R programlama dilinde Kuhn ve diğerleri tarafından 2015 yılında geliştirilen C5.0 kütüphanesi kullanılarak algoritma uygulanmıştır. Program kod bilgileri ve detaylı sonuç bilgisi Ek 7'de gösterilmektedir.

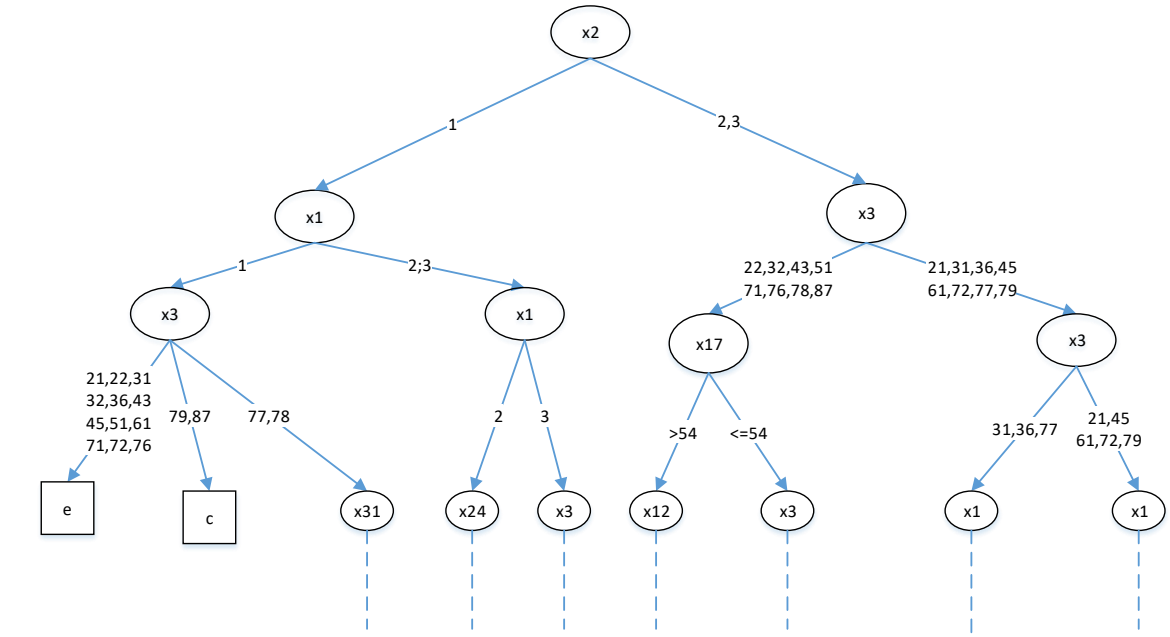
Şekil 16: 5+ veri seti - algoritma çıktısı karar ağacı



Şekil 17: SES AB veri seti - algoritma çıktısı karar ağacı



Şekil 18: 20+ABC1 veri seti - algoritma çıktısı karar ağacı



Uygulamada öncelikle veri programa tanıtılmış kategorik değişkenlerin bilgisi verilmiş, veri seti eğitim ve test seti olarak rastgele ayrılmıştır. Eğitim seti kullanılarak karar ağacı

algoritması ile öğrenme modeli kurulmuştur. Test setindeki sınıf değişkenleri temizlenerek algoritmanın tahmin yapması sağlanmıştır. Öznitelik boyutunun yüksek olmasından dolayı karar ağacı görselleri tamamıyla paylaşılamamış Şekil 17, Şekil 18 ve Şekil 19'daki gibi küçük parçaları sunulmuştur.

Tablo 14: 3 veri seti için de karşılaştırma matrisleri

5+

Tahmin \ Gerçek	a	b	c	d	e
a	97	25	2	0	1
b	13	325	29	5	1
c	0	15	28	4	1
d	0	0	1	8	10
e	0	0	2	12	36

SES AB

Tahmin \ Gerçek	a	b	c	d	e
a	84	28	0	1	1
b	16	303	47	0	1
c	0	34	38	3	0
d	0	0	2	5	2
e	0	0	5	3	42

20+ ABC1

Tahmin \ Gerçek	a	b	c	d	e
a	77	19	1	1	0
b	14	298	35	5	0
c	0	38	56	2	0
d	0	0	8	6	0
e	0	0	2	6	47

Tablo 14'de gösterilen karşılaştırma matrisinde ilgili test setlerindeki verilerin karar ağacı algoritması ile tahmin edildiği değerler ve gerçek sınıf değerleri gösterilmektedir. "5+" olarak adlandırılan 5 yaş üstü tüm bireylere göre belirlenen reyting sınıflarının kullanıldığı veri seti ile karar ağacı algoritması uygulandığında;

- test setinde gerçek sınıfı "a" olan 110 verinin 97 tanesini "a" sınıfına,
- test setinde gerçek sınıfı "b" olan 365 verinin 325 tanesini "b" sınıfına,
- test setinde gerçek sınıfı "c" olan 62 verinin 28 tanesini "c" sınıfına,

- test setinde gerçek sınıfı “d” olan 29 verinin 8 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 49 verinin 36 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

“SES AB” olarak adlandırılan sosyoekonomik statü gruplarından A ve B’ye dâhil olan çoğunluğu lisans ve lisansüstü düzeyde eğitim görmüş bireylere göre belirlenen reyting sınıflarının kullanıldığı veri seti ile karar ağacı algoritması uygulandığında,

- test setinde gerçek sınıfı “a” olan 100 verinin 84 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 365 verinin 303 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 92 verinin 38 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 12 verinin 5 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 46 verinin 42 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

“20+ ABC1” olarak adlandırılan sosyoekonomik statü gruplarından A, B ve C1’e dâhil olan 20 yaş üstü ve çoğunlukla eğitim seviyeleri lisansüstü ve lise arasında değişen eğitim seviyeli bireylere göre belirlenen reyting sınıflarının kullanıldığı veri seti ile karar ağacı algoritması uygulandığında,

- test setinde gerçek sınıfı “a” olan 91 verinin 77 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 355 verinin 298 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 102 verinin 56 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 20 verinin 6 tanesini “d” sınıfına,

test setinde gerçek sınıfı “e” olan 47 verinin 47 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

Sosyal medya kanalı olarak seçilen Twitter’dan derlenen seyirci görüşlerinin nitelik ve nicelik olarak analizleri ve programların kendine has özellikleri ile gelecek programın reyting değerinin tahminlenmesi amacıyla kurulan karar ağacı modeli doğruluk değerleri gösteriyor ki ilgili model her üç veri seti için de %75’in üzerinde bir doğruluk oranına

sahiptir. 5 yaş üstü tüm bireylere göre hesaplanan reytingin sınıf değişkeni olduğu veri setinde tahmin modelin doğruluk oranı %80.32 iken SES AB grubunda %76.74, 20 yaş üstü ABC1 grubunda ise %78,69 olarak hesaplanmıştır. Böylece bu öznitelikler/tahminleyiciler kullanarak her üç reyting tipi için de yapılacak olan gelecek planında doğru sonuca ulaşma oranı yüksek olacaktır. Sosyoekonomik gruplara göre doğruluk oranının farklı çıkması ve 5 yaş üstü bireyleri oluşturan grubunkinin daha yüksek çıkması, örneklemi oluşturan kitlenin yani televizyon izleyicisi olup Twitter hesabı olan kişilerin her yaştan her eğitim ve gelir seviyesinden olabileceğinin göstergesidir.

Birçok karar ağacı algoritması gibi C5.0'da bölme işlemlerini yaparken kullandığı öznitelikleri önem sırasına göre göstermektedir. 5+ veri setinde algoritmanın ortaya koyduğu önem arz eden özniteliklerden ilk beşi program adı ile pozitif duygulu atılan tweetler, program türü, hitap ettiği kitle bilgisi, program süresince atılan toplam tweet sayısı ve hangi gün yayınlandığıdır. SES AB veri setinde algoritmanın bölme işleminde kullandığı önemli öznitelikler, programın hitap ettiği kitle, programın türü, program başlamadan 2 saat önceden ve bittikten 2 saat sonraya kadar olan aralıkta atılan toplam tweet sayısı, kanal adı ile atılan pozitif duygulu tweet sayısıdır. 20+ ABC1 veri setindeki önemli bölen öznitelikler ise kanal adı ile atılan pozitif duygulu tweetlerin retweet sayısı, programın yayınlandığı gün, kanal adı ile atılan toplam tweet sayısı, kanal adı ile atılan pozitif duygulu tweet sayısı ve program başlamadan 2 saat önceden ve bittikten 2 saat sonraya kadar olan aralıkta atılan toplam tweet sayısıdır. Böylece sosyal medyadan elde edilen özniteliklerin tahmin algoritmasında önem taşıdığı, bir programın gelecek yayındaki reyting değeri program hakkında yayın süresince ve başlamadan önce ve bittikten sonra atılan tweet sayıları ve bunların pozitif duygulu olanlarının sayıları ile tahmin edilebileceği ortaya konulmuştur.

C5.0 algoritmasının bir özelliği de kural modelleri geliştirmesidir. Bu kurallar özniteliklerin belli değerlerine göre atanacakları sınıfları belirtmektedir. Buna göre 5+ veri seti için algoritmanın oluşturduğu kurallardan bazıları şöyledir;

- Eğer x3 = 79 ve x6 = 1 ve x27 <= 593 ise "a" sınıfı.

- Eğer x1 = 3 ve x2 = 1 ve x4 = 1 ise “b” sınıfı.
- Eğer x2 = 2 ve x3-{71,76} ve x6-{5,7}
  x17 <= 0 ve x22 > 54 ve x27 > 3 ise “c” sınıfı.
- Eğer x2 = 1 ve x6 = 7 ve x22 > 4 ve x25 <= 44 ise “d” sınıfı.
- Eğer x1 = 1 ve x2 = 3 ve x3 = 76
  x6-{1,5,7} ve x9 > 18 ve x17 <= 0 ise “e” sınıfı.

SES AB veri seti için algoritmanın oluşturduğu kurallardan bazıları şu şekildedir;

- Eğer x3 = 45 ve 5 < x13 <= 38 ve x16 > 0 ise “a” sınıfı.
- Eğer x3-{71,76} ve x6 = 5 ve 0 < x27 <= 4 ise “b” sınıfı.
- Eğer x1 = 2 ve x3 = 76 ve x4 = 1 ve x6-{4,7}
  x13 > 11 ve x17 <= 2 ise “c” sınıfı.
- Eğer x2 = 1 ve x3 = 78 ve x6 = 4 ve x14 <= 4 ise “d” sınıfı.
- Eğer x1 = 1 ve x3 = 76 ve x9 >19 ise “e” sınıfı.

20+ ABC1 veri seti için algoritmanın oluşturduğu kurallardan bazıları şu şekildedir;

- Eğer x2-{2,3} ve x3-{45,77} ve x6 = 4 ve x27> 4 ise “a” sınıfı.
- Eğer x1 = 3 ve x3 = 72 ise “b” sınıfı.
- Eğer x1 = 2 ve x3 = 76 ve x6 = 4
  x13 > 11 ve x23 <= 6 ise “c” sınıfı.
- Eğer x2 = 3 ve x8 <= 2916
  x9 > 206 ve x13 > 54 ise “d” sınıfı.
- Eğer x2 = 3 ve x3 = 76 ve x8 > 2916 ise “e” sınıfı.

Bu aşamada veri setlerinden anlamlı örüntüler ve sınıflama kuralları keşfedebilmek için uygulanması planlanan ilk veri madenciliği yöntemi olan karar ağacı yöntemi C5.0 algoritması uygulanmıştır. Tüm veri setleri için %75’in üzerinde doğruluk oranı veren algoritma ortaya koyduğu kurallar ile de karar vericilere yol gösterebilmektedir.

5.5.2.2. Destek Vektör Makineleri

Verilerin optimal olarak sınıflandırılması amacıyla çok boyutlu uzayda hiperdüzlem oluşturan destek vektör makineleri girdi ve çıktı eşleme fonksiyonları oluşturulan bir denetimli öğrenme yöntemidir. Bu eşleme fonksiyonu sınıflandırma ve regresyon amaçlı kullanılabilmektedir. Sınıflandırma problemlerinde çoğunlukla girdilerin orijinal giriş uzayından daha kolay ayrışabilmesi için yüksek boyutlu nitelik uzayına dönüştürülmesini sağlayan doğrusal olmayan çekirdek (kernel) fonksiyonları kullanılmaktadır.

R programlama dilinde destek vektör makineleri algoritması için geliştirilen çeşitli kütüphaneler bulunmaktadır. e1071 adlı kütüphane de bunlardan birisidir. İçerisinde çeşitli veri analizi algoritmalarını barındıran kütüphane "SVM" fonksiyonu ile gelişmiş bir destek vektör makineleri algoritması uygulamasına olanak sağlar. Öte yandan sağladığı "tune" fonksiyonu ile çeşitli kernel fonksiyonlarının parametre alternatiflerini tek bir kod ile deneyerek en düşük hata oranını veren uygun parametre değerlerini sunmaktadır. Bu tez çalışmasında e1071 kütüphanesinin 2017 yılında güncellenmiş 1.6-8 sürümü kullanılmıştır, kod ve detaylı sonuç çıktıları Ek 8'de paylaşılmaktadır. 27 öznitelik bir sonuç değişkeni var olan veri setindeki 6 kategorik özniteliğin kukla değişkene dönüştürülmesi ile toplam öznitelik sayısı 55'e çıkmıştır. Bu 55 öznitelik 1 sınıf değişkenli veri setleri diğer algoritmalarda olduğu gibi %80'i eğitim %20'si test seti olacak şekilde rasgele ikiye ayrılmıştır ve böylece tahmin değerleri ile gerçek değerler karşılaştırılarak tahmin başarısı ortaya çıkarılmıştır.

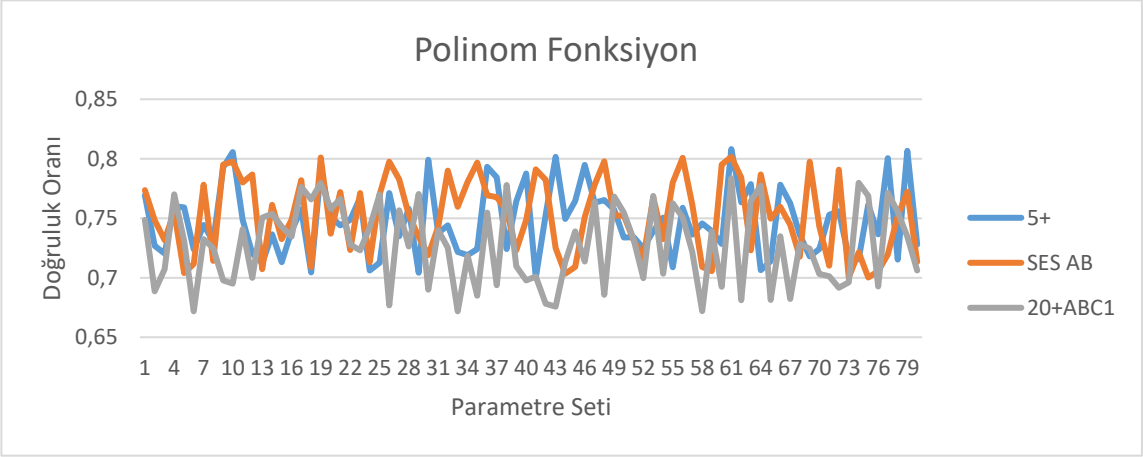
Tablo 15: DVM algoritması kernel fonksiyonlarında kullanılan parametreler

Parameters	Polinom Fonksiyon	Radyal Temelli Fonksiyon	Doğrusal Fonksiyon
Derece (d)	1,2,3,4	-	-
Gamma (γ)	0.001,0.01,….10	0.001,0.01,….10	-
Ceza katsayısı (c)	1,10,100,1000	1,10,100,1000	-

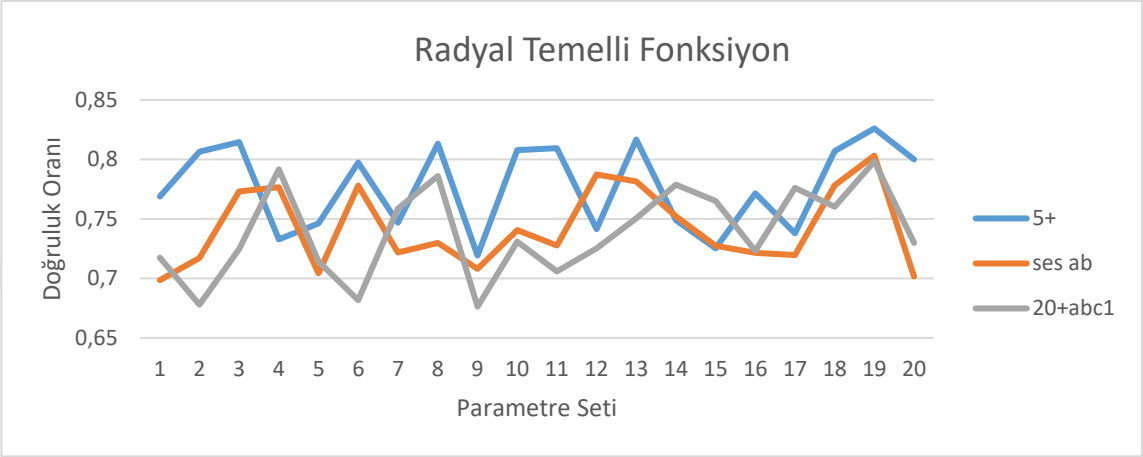
Literatürde sıklıkla yer bulan kernel fonksiyonları; doğrusal, polinom ve radyal temelli fonksiyonlar kullanılarak üç farklı destek vektör makineleri algoritması kullanılmıştır. Üç

fonksiyonun da kendine has parametre değerleri vardır. Algoritma, her parametre değeri için farklı sonuç verdiğinden en iyi sonucu veren parametre setinin belirlenip tahmin modelinde bu değerlerin kullanılması önemlidir. Bu amaçla Tablo 15'de belirtilen parametrelerle üç veri seti için de destek vektör makineleri algoritması uygulanmıştır. Dolayısı ile her veri seti için 80 polinom fonksiyonlu DVM, 20 radyal temelli fonksiyon ile DVM ve 1 doğrusal fonksiyonlu DVM olmak üzere toplam 101 model denenmiş ve en uygun sonucu veren parametreler belirlenmiştir. Bu amaçla tüm parametrelerin girdi olarak verildiği tune fonksiyonu kullanılarak, en iyi ayırma özelliğine sahip en düşük hata oranını veren parametre seti elde edilmiştir (Şekil 19- Şekil 20). Bu bilgiler ışığında DVM algoritmaları uygulanırken ilgili parametreler kullanılmış ve eğitilen model sayesinde test seti için tahmin yapılarak doğruluk oranları hesaplanmış ve kernel fonksiyonlar bazında karşılaştırma yapılmıştır.

Şekil 19: Polinom fonksiyonda kullanılan parametre setlerine göre elde edilen doğruluk oranları



Şekil 20: Radyal temelli fonksiyonda kullanılan parametre setlerine göre elde edilen doğruluk oranları



Tüm veri setleri için uygulanan DVM algoritmasının kernel fonksiyonlarındaki en uygun parametre setleri ve doğruluk oranları Tablo 16'da gösterilmektedir. Doğrusal fonksiyonun herhangi bir parametresi yoktur, bu nedenle eğitim setine öğrenme modeli doğrudan uygulanmış ve buna göre test seti verilerine tahmin yapılmıştır. Polinom ve radyal temelli fonksiyonlar ise tüm parametre setlerine göre denenmiştir.

Tablo 16: Fonksiyonlara göre en uygun parametreler ve doğruluk oranları

		5+		
		Doğrusal Fonksiyon	Polinom Fonksiyon	Radyal Temelli Fonksiyon
En Uygun Parametreler	Derece	-	3	-
	Gamma	-	0.01	0.01
	Ceza Katsayısı	-	100	10
Doğruluk Oranı		81.60%	80.81%	82.60%

		SES AB		
		Doğrusal Fonksiyon	Polinom Fonksiyon	Radyal Temelli Fonksiyon
En Uygun Parametreler	Derece	-	3	-
	Gamma	-	0.01	0.01

	Ceza Katsayısı	-	100	10
Doğruluk Oranı		76.42%	80.16%	80.32%

		20+ABC1		
		Doğrusal Fonksiyon	Polinom Fonksiyon	Radyal Temelli Fonksiyon
En Uygun Parametreler	Derece	-	3	-
	Gamma	-	0.01	0.01
	Ceza Katsayısı	-	100	10
Doğruluk Oranı		78.54%	78.37%	79.84%

Her üç veri seti içinde polinom fonksiyonda derecenin 3, gammanın 0.01 ve ceza katsayısının 100 olmasının uygun sonuç verdiği anlaşılmıştır. Radyal temelli fonksiyonda ise gamma 0.01 ve ceza katsayısı 10 olduğunda en uygun tahmin başarısı elde edilmiştir. Tüm veri setlerine, üç farklı kernel fonksiyonuyla uygulanan DVM algoritmasında %78'in üzerinde tahmin doğruluk oranına ulaşıldığı görülmektedir.

Tablo 17: 5+ veri seti için karşılaştırma matrisleri

Doğrusal Fonksiyon

Tahmin \ Gerçek	a	b	c	d	e
a	98	30	1	0	0
b	14	334	34	5	0
c	0	7	28	2	0
d	0	0	1	6	6
e	0	0	4	9	36

Polinom Fonksiyon

Tahmin \ Gerçek	a	b	c	d	e
a	95	22	1	0	0
b	17	339	39	6	2
c	0	10	26	3	1
d	0	0	0	4	6
e	0	0	2	9	33

Radyal Temelli Fonksiyon

Tahmin \ Gerçek	a	b	c	d	e
a	97	26	1	0	0
b	14	337	36	6	0
c	0	8	27	1	0

	d	0	0	0	6	1
	e	1	0	4	9	41

Karşılaştırma matrisinde (Tablo 17) de görüldüğü gibi 5+ veri setine doğrusal fonksiyon ile destek vektör makineleri algoritması uygulandığında;

- test setinde gerçek sınıfı “a” olan 112 verinin 98 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 371 verinin 334 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 68 verinin 28 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 22 verinin 6 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 42 verinin 36 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

%81.60 doğruluk oranı sayesinde sosyal medya unsurları ile gelecek tahmini yapılabilmesi amacıyla kurulan tahmin modelinin yüksek başarıya sahip olduğu söylenebilmektedir.

5+ veri setine polinom fonksiyon destek vektör makineleri algoritması uygulanması amacıyla derece değeri 3, gamma değeri 0.01 ve ceza katsayısı 100 olarak belirlenen parametre seti seçilmiştir. Algoritma;

- test setinde gerçek sınıfı “a” olan 112 verinin 95 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 371 verinin 339 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 68 verinin 26 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 22 verinin 4 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 42 verinin 33 tanesini “e” sınıfına atamış yani doğru tahminleyebilmiştir.

Bu tahminlerle %80.81 oranında doğruluk sunduğu ortaya çıkan destek vektör makineleri algoritması, ilgili öznitelikler ile kaliteli bir gelecek dönem reyting tahmini yapılabileceğini göstermiştir.

5+ veri setine radyal temelli fonksiyon gamma değeri 0.01 ve ceza katsayısı 10 iken destek vektör makineleri algoritması uygulandığında;

- test setinde gerçek sınıfı “a” olan 112 verinin 97 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 371 verinin 337 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 68 verinin 27 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 22 verinin 6 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 42 verinin 41 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

5 yaş üstü bireylere göre ölçülen reyting değerlerinin sınıf değişkeni olarak kullanıldığı veri setinde kurulan tahmin modelinde en yüksek doğruluk oranını %82.60 ile radyal temelli fonksiyonun kullanıldığı destek vektör makineleri algoritması vermiştir.

SES AB olarak adlandırılan sosyoekonomik grubun yer aldığı veri seti için üç farklı kernel fonksiyonu ile destek vektör makineleri algoritması uygulanmıştır ve elde edilen karşılaştırma matrisleri Tablo 18’de gösterilmektedir.

Tablo 18: SES AB veri seti için karşılaştırma matrisleri

Doğrusal Fonksiyon

		Gerçek				
		a	b	c	d	e
Tahmin	a	87	20	3	1	0
	b	20	315	66	3	1
	c	2	4	6	3	0
	d	0	0	5	6	2
	e	0	0	4	11	56

Polinom Fonksiyon

		Gerçek				
		a	b	c	d	e
Tahmin	a	89	24	2	1	0
	b	15	306	46	2	0
	c	4	8	33	2	5
	d	0	1	1	12	1
	e	1	0	2	7	53

Radyal Temelli Fonksiyon

		Gerçek				
		a	b	c	d	e
Tahmin	a	89	19	2	1	0
	b	19	312	46	1	0
	c	1	7	30	2	3

	d	0	1	3	10	3
	e	0	0	3	10	53

Karşılaştırma matrisinde de görüldüğü gibi SES AB veri setine doğrusal fonksiyon ile destek vektör makineleri algoritması uygulandığında;

- test setinde gerçek sınıfı “a” olan 109 verinin 87 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 339 verinin 315 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 84 verinin 6 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 24 verinin 6 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 59 verinin 56 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

%76.42 doğruluk oranı sayesinde sosyal medyadan derlenen veriler ile gelecek tahmini yapılması amaçlanan tahmin modelinin yüksek başarıya sahip olduğu söylenebilmektedir.

SES AB veri seti ile destek vektör makineleri algoritmasında polinom fonksiyon kullanılarak tahmin modeli kurulmuştur. Algoritmada derece değeri 3, gamma değeri 0.01 ve ceza katsayısı 100 olarak belirlenen parametre seti seçilmiştir ve uygulanmıştır. Algoritma;

- test setinde gerçek sınıfı “a” olan 109 verinin 89 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 339 verinin 306 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 84 verinin 33 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 24 verinin 12 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 59 verinin 53 tanesini “e” sınıfına atamış yani doğru tahminleyebilmiştir.

Böylece %80.16’lık doğruluk oranına sahip bir tahmin modelinin geliştirildiği ve bunun programa has özellikler ve seyirci görüşlerinin Twitter ile nitelik ve nicelik değerleri ile gelecek bölüm reyting tahmininde kullanılabilecek bir model olduğu ortaya konulmuştur.

SES AB veri setine radyal temelli fonksiyon gamma değeri 0.01 ve ceza katsayısı 10 iken destek vektör makineleri algoritması uygulandığında ise ;

- test setinde gerçek sınıfı “a” olan 109 verinin 89 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 339 verinin 312 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 84 verinin 30 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 24 verinin 10 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 59 verinin 53 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

Sosyoekonomik gruplardan A ve B grubundaki bireylere göre ölçülen reyting değerlerinin sınıf değişkeni olarak kullanıldığı veri setinde kurulan tahmin modelinde en yüksek doğruluk oranını %80.32 ile radyal temelli fonksiyonun kullanıldığı destek vektör makineleri algoritması vermiştir.

20+ABC1 veri seti 20 yaş üzeri adlandırılan sosyoekonomik grubu A, B ve C1 olan bireylere göre ölçülen reyting değerlerini içermektedir. 20+ABC1 veri seti için de üç kernel fonksiyonun ayrı ayrı kullanıldığı üç destek vektör makineleri algoritması uygulanmıştır ve elde edilen karşılaştırma matrisleri Tablo 19’da gösterilmektedir.

Tablo 19: 20+ABC1 veri seti için karşılaştırma matrisleri

Doğrusal Fonksiyon

		Gerçek				
		a	b	c	d	e
Tahmin	a	84	21	5	0	0
	b	14	312	72	4	0
	c	0	5	19	3	0
	d	0	0	0	10	0
	e	0	0	2	6	58

Polinom Fonksiyon

		Gerçek				
		a	b	c	d	e
Tahmin	a	87	29	1	0	0
	b	10	296	56	6	0
	c	0	12	39	1	6
	d	1	1	0	12	4
	e	0	0	2	4	48

Radyal Temelli Fonksiyon

	Gerçek				
	a	b	c	d	e

Tahmin	a	87	25	0	0	0
	b	11	301	60	4	0
	c	0	12	35	3	1
	d	0	0	1	11	0
	e	0	0	2	5	57

Yukarıdaki karşılaştırma matrisinde görüldüğü gibi 20+ABC1 veri setine doğrusal fonksiyon kullanılarak destek vektör makineleri algoritması uygulandığında;

- test setinde gerçek sınıfı “a” olan 98 verinin 84 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 338 verinin 312 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 98 verinin 19 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 23 verinin 10 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 58 verinin 58 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

Böylece %78.54 doğruluk oranı sayesinde sosyal medyadaki seyirci görüşleri ve programların özellikleri ile gelecek tahmini yapılması amaçlanan tahmin modelinin yüksek başarıya sahip olduğu söylenebilmektedir.

20+ABC1 veri seti ile destek vektör makineleri algoritmasında polinom fonksiyon kullanılarak kurulan tahmin modelinde derece değeri 3, gamma değeri 0.01 ve ceza katsayısı 100 olarak belirlenen parametre seti seçilmiştir. Uygulama ile algoritma;

- test setinde gerçek sınıfı “a” olan 98 verinin 87 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 338 verinin 296 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 98 verinin 39 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 23 verinin 12 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 58 verinin 48 tanesini “e” sınıfına atamış yani doğru tahminleyebilmiştir.

%78.37'lik doğruluk oranına sahip bir tahmin modelinin geliştirildiği ve bunun program özellikleri ve seyirci görüşlerinin Twitter'dan derlenerek analiz edilmesi ile gelecek bölüm reyting tahmininde kullanılabilecek bir model olduğu ortaya konulmuştur.

20+ABC1 veri setine radyal temelli fonksiyon gamma değeri 0.01 ve ceza katsayısı 10 iken destek vektör makineleri algoritması uygulandığında;

- test setinde gerçek sınıfı "a" olan 98 verinin 87 tanesini "a" sınıfına,
- test setinde gerçek sınıfı "b" olan 338 verinin 301 tanesini "b" sınıfına,
- test setinde gerçek sınıfı "c" olan 98 verinin 35 tanesini "c" sınıfına,
- test setinde gerçek sınıfı "d" olan 23 verinin 11 tanesini "d" sınıfına,
- test setinde gerçek sınıfı "e" olan 58 verinin 57 tanesini "e" sınıfına atanmış yani doğru tahminlenebilmiştir.

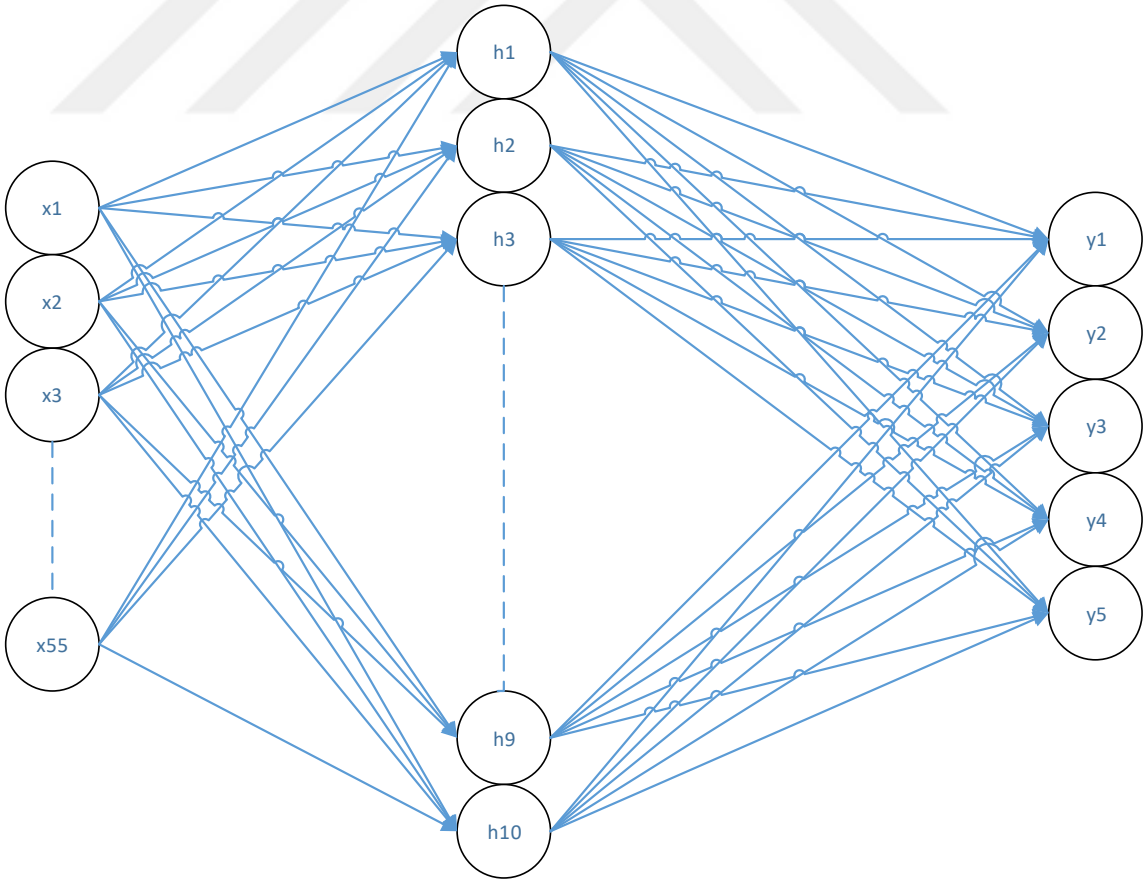
20 yaş üzeri ve sosyoekonomik olarak A, B ve C1 grubunda yer alan bireylere göre ölçülen reyting değerlerinin sınıf değişkeni olarak kullanıldığı veri setinde kurulan tahmin modelinde en yüksek doğruluk oranını %79.84 ile radyal temelli fonksiyonun kullanıldığı destek vektör makineleri algoritmasının verdiği gözlenmiştir.

Her üç veri seti içinde kernel fonksiyon çeşitleri ve onların parametre çeşitleri için destek vektör makineleri algoritması ayrı ayrı uygulanmıştır. Buna göre destek vektör makineleri algoritması ile belirlenen en uygun parametre setleri kullanılarak eğitim veri seti ile model eğitilmiştir. Ardından test setindeki nesnelerin sınıf değerlerinin tahminlenmiş ve modelin tahmin performansının belirlenebilmesi için karşılaştırma matrisleri oluşturulmuş, performans değerlendirme ölçütlerinden biri olan doğruluk oranları hesaplanmıştır. Bu sayede her veri seti için en yüksek doğruluk oranını veren kernel fonksiyonun radyal temelli fonksiyon olduğu ortaya çıkarılmıştır. Böylece diğer sınıflandırma algoritmaları ile kıyaslanırken radyal temelli fonksiyonlu DVM kullanılabilecektir.

5.5.2.3. Yapay Sinir Ağları

Yapar sinir ağları adından da anlaşılabileceği üzere, sinir hücreleri yani nöronların insan beynindeki bilişsel davranış sürecindeki aktivitelerinden esinlenilerek geliştirilmiştir. Nöronların öğrenme yapısını taklit eden algoritma, girdi ve çıktı değişkenleri arasındaki bağlantılar vasıtasıyla modelin eğitilmesini ve yeni bilgiler keşfedilebilmesini sağlamaktadır. Çalışmada, geri besleme ya da bir katmanın atlanması söz konusu olmayan öte yandan hata sinyallerine göre bağlantı ağırlıkları güncelleyen geriye yayılım ağları kullanılmıştır. Ağ yapısı kurgulanırken Şekil 21'de görüldüğü gibi tek gizli katmanlı yapı seçilmiştir.

Şekil 21:Yapay sinir ağı



Bu amaçla The Stuttgart Neural Network Simulator (SNNS) adlı pek çok sinir ağları uygulaması barındıran kütüphanenin R programlama dili için geliştirilmiş olanı RSNNS kullanılmıştır. Kütüphane içerisinde çeşitli sinir ağları barındırmaktadır. Bu çalışma için geriye yayılım ağları kullanılacağından "mlp" fonksiyonu seçilmiştir (Ek 9).

Tablo 20: ANN modelinde kullanılan parametreler

Parameterler	Değerler
Gizli katman sayısı (n)	10,20,30,……,...,90,100
Momentum (mc)	0.1, 0.2,……..…..0.8, 0.9
Öğrenme katsayısı (l)	0.01,0.05,0.1……..,.0.45,0.5
Tekrar sayısı (ep)	100,500,1000,…4500, 5000

Mlp fonksiyonu, öğrenme fonksiyonu olarak geriye yayılım ağlarını kullanmaktadır. Fonksiyon için belirlenmesi gereken parametrelerden bazıları ise gizli katman sayısı ve her katmanda yer alacak gizli nöron sayısı, öğrenme katsayısı, momentum değeri ve tekrar sayısıdır. Bu değerlerin her veri seti için en iyi tahmin performansını verdiği bir parametre seti ya da ideal bir aralığı olmadığından parametre optimizasyonu yapılabilmesi için Tablo 20'de yer alan 10x9x11x11 adet parametre seti denenerek algoritma uygulanmıştır. Dolayısı ile üç veri seti için de 10890 adet yapay sinir ağları algoritması denemesi yapılarak her veri seti için en uygun sonucu veren set belirlenmiştir. Öte yandan destek vektör makinelerindeki gibi kategorik değişkenlerin kukla değişkenlere dönüştürüldüğü veri seti kullanıldığından 55 öznitelik ve bir çıktısı bulunan 3 farklı veri seti üzerinde algoritma uygulanmış ve tahmin performansları elde edilmiştir.

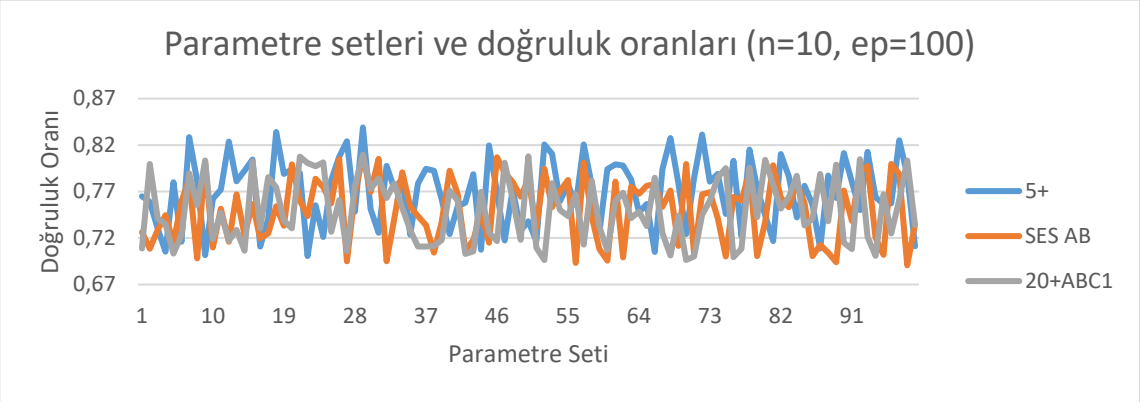
Tablo 21: En uygun parametreler

	5+	SES AB	20+ABC1
Gizli katman sayısı (n)	10	10	10
Momentum (mc)	0.3	0.3	0.3
Öğrenme katsayısı (l)	0.01	0.01	0.01
Tekrar sayısı (ep)	100	100	100
Doğruluk Oranı	83.90%	80.81%	80.97%

Tablo 21 en uygun sonucu veren yani tüm parametre setleri içerisinde en yüksek tahmin performansını, doğruluk oranını sağlayan parametre seti ve tahmin performansı hakkında

bilgi vermektedir. Şekil 22'de gizli katman sayısı 10, tekrar sayısı 100 iken öğrenme katsayısı ve momentum parametrelerindeki değişim ile doğruluk oranları gösterilmiştir.

Şekil 22: Geriye Yayılım algoritmasında kullanılan parametre setlerine göre elde edilen doğruluk oranları



En uygun parametre setlerinin belirlenmesi kurulacak yapay sinir ağları modelinde kullanılması en yüksek doğruluk oranını sağlayacak ağ yapısına ulaşılmasını sağlamaktadır. Her üç veri seti için de 55 girdi nöronu, 1 gizli katmanı, gizli katmanda 10 nöronu ve bir adet çıktı nöronu bulunan geriye yayılım ağı kurulmuştur. Bu ağlar öncelikle tüm veri setinin %80'i olarak ayrılan eğitim seti ile eğitilmiş ardından %20'lik test seti vasıtasıyla tahmin başarısı ölçülmüştür. Ağların her üç veri setinden ayrılan test setlerinde tahmin ettiği çıktı değerleri ile gerçek çıktı değerlerinin karşılaştırıldığı karşılaştırma matrisleri Tablo 22'de yer almaktadır.

Tablo 22: 3 veri seti için de karşılaştırma matrisleri

5+

Tahmin \ Gerçek	a	b	c	d	e
a	113	15	0	0	0
b	19	330	4	0	0
c	2	34	21	1	4
d	1	3	2	8	10
e	0	0	1	3	44

SES AB

Tahmin \ Gerçek	a	b	c	d	e
a	96	20	1	0	0
b	20	316	9	0	0
c	0	46	29	1	4
d	0	2	3	4	11
e	0	1	0	0	52

20+ABC1

		Gerçek				
		a	b	c	d	e
Tahmin	a	89	19	0	0	0
	b	20	294	16	0	0
	c	0	42	48	0	4
	d	1	2	3	6	8
	e	1	0	0	1	61

“5+” olarak adlandırılan 5 yaş üstü tüm bireylere göre belirlenen reyting sınıflarının kullanıldığı veri seti ile yapay sinir ağı algoritması uygulandığında;

- test setinde gerçek sınıfı “a” olan 135 verinin 113 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 382 verinin 330 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 28 verinin 21 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 12 verinin 8 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 58 verinin 44 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

“SES AB” olarak adlandırılan sosyoekonomik statü gruplarından A ve B’ye dâhil olan bireylere göre belirlenen reyting sınıflarının kullanıldığı veri seti ile yapay sinir ağı algoritması uygulandığında,

- test setinde gerçek sınıfı “a” olan 116 verinin 96 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 385 verinin 316 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 42 verinin 29 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 5 verinin 4 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 67 verinin 52 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

“20+ ABC1” olarak adlandırılan sosyoekonomik statü gruplarından A, B ve C1’e dâhil olan 20 yaş üstü bireylere göre belirlenen reyting sınıflarının kullanıldığı veri seti ile yapay sinir ağı algoritması uygulandığında,

- test setinde gerçek sınıfı “a” olan 111 verinin 89 tanesini “a” sınıfına,
- test setinde gerçek sınıfı “b” olan 357 verinin 294 tanesini “b” sınıfına,
- test setinde gerçek sınıfı “c” olan 67 verinin 48 tanesini “c” sınıfına,
- test setinde gerçek sınıfı “d” olan 7 verinin 6 tanesini “d” sınıfına,
- test setinde gerçek sınıfı “e” olan 73 verinin 61 tanesini “e” sınıfına atanmış yani doğru tahminlenebilmiştir.

Sosyal medyanın kullanıcı görüşlerinin direkt olarak elde edebilme olanağı sunması işletmeler ve akademisyenler için yeni bir araştırma alanı yaramıştır. Çalışmada amaçlandığı gibi sosyal medya kanallarının gerçek satış rakamlarına etkisinin ölçülebilmesi ya da tahmin etme imkânı vermesi işletmeler için değer yaratmaktadır. Yapay sinir ağları algoritması ile sosyal medyadan elde edilen seyirci görüşlerinin gelecek program reyting tahmini yapılması 5 yaş üstü bireylere göre ölçülen reyting veri setinde %83.90’lık doğruluk sağlamıştır. Böylece yeni bir veri geldiğinde bunun doğru reyting sınıfına atılma olasılığı %83.90’dır. Sosyoekonomik gruplara göre belirlenen reyting veri setlerinden SES AB veri setinde ise algoritma %80.81 oranında doğruluk sunmaktadır. 20 yaş üzeri ve sosyoekonomik olarak A B ve C1 sınıfında olan bireylere göre belirlenen reytingleri içeren veri setinde ise algoritma %80.97’lik doğruluk oranı vermiştir. Yapay sinir ağları algoritması ile üç veri seti için de yüksek doğruluk oranları elde edilmesi ilgili algoritmanın oluşturulan tahmin modelinde kullanılmaya uygun olduğunu göstermektedir.

Yapay sinir ağları algoritmasının yorumlanarak değer yaratılmasına imkân sunan ağırlık sonuçları ağın boyutu çok büyük olduğu için paylaşılamamış Ek 10’da matrisler halinde sunulmuştur. Sinir ağları arasındaki ilişkilerin gücünü gösteren bu ağlar girdi nöronunun çıktıyı ne kadar etkilediğini göstermektedir. Her üç veri seti için de ağırlık matrisleri

incelenmiş ve sınıf değerlerini en yüksek derecede etkileyen gizli nöronlar ve ilgili gizli nöronları en yüksek derecede etkileyen ilk üç girdi nöronu Tablo 23'de özetlenmiştir.

Tablo 23: Ağırlık Matrisleri

5+		
Çıktı Nöronu	En Yüksek Ağırlıklı Gizli Nöron	En Yüksek Ağırlıklı İlk Üç Girdi Nöronu
a	h3	x3.77/x3.72/x5.3
b	h2	x3.71/x3.22/x5.3
c	h8	x3.79/x3.71/x9
d	h5	x6.5/x3.79/x9
e	h6	x3.76/x5.3/x27

SES AB		
Çıktı Nöronu	En Yüksek Ağırlıklı Gizli Nöron	En Yüksek Ağırlıklı İlk Üç Girdi Nöronu
a	h10	x3.72/x3.77/x5.3
b	h2	x6.3/x3.22/x3.43
c	h6	x3.79/x3.72/x4.0
d	h3	x3.87/x5.1/x1.1
e	h4	x3.76/x1.3/x7

20+ABC1		
Çıktı Nöronu	En Yüksek Ağırlıklı Gizli Nöron	En Yüksek Ağırlıklı İlk Üç Girdi Nöronu
a	h6	x3.72/x4.0/x6.5
b	h5	x3.71/x3.43/x3.22
c	h3	x3.22/x3.79/x3.71
d	h2	x3.76/x25/x24
e	h7	x3.76/x3.43/x27

Yukarıdaki tabloda her çıktı nöronunu en yüksek pozitif ağırlıkla etkileyen gizli nöron ve bu gizli nöronu en yüksek pozitif ağırlıkla etkileyen ilk üç girdi nöronu gösterilmektedir. 5+ veri seti için a sınıfını en çok etkileyen gizli nöronun h3 olduğu ve onu en yüksek ağırlık değerleri ile etkileyen girdi değişkenlerinin ise yayın türü (77-spor karşılaşmaları, 72-skeç, tiyatro vb) ve ünlülük seviyesi (3-düşük) olduğu görülmüştür. En yüksek reyting değerlerinin sınıf değeri olan e ise en çok h6 gizli nöronundan etkilenmektedir. Bu gizli

nöronu en yüksek pozitif ağırlıkla etkileyen ilk üç girdi nöronu ise; yayın türü (76-yarışma), ünlülük seviyesi (3-düşük) ve program başlamadan 2 saat önce ve başladıktan 2 saat sonra program adında atılan toplam tweet sayısıdır. SES AB veri setinde ise e sınıf nöronunu en yüksek ağırlıkla etkileyen gizli nöron h4 olarak belirlenmiştir. Yayın türü (76-yarışma), hitap ettiği kite (3-13+) ve program süresince atılan toplam tweet sayısı da h4 gizli nöronunu etkileyen önemli girdi nöronlarıdır. 20+ABC1 veri setinde ise e sınıfında önemli bir ağırlığa sahip olan gizli nöron h7'dir. Bu nöron yayın türü (76-yarışma, 43-bilgi beceri yemek programları) ve program başlamadan 2 saat önce ve başladıktan 2 saat sonra program adında atılan toplam tweet girdi nöronlarından etkilenmektedir. Bu ağırlıklar ile tahmin modelinde programa has özellikler ve sosyal medya verilerinin önemli etkilere sahip olduğu ortaya çıkarılmıştır.

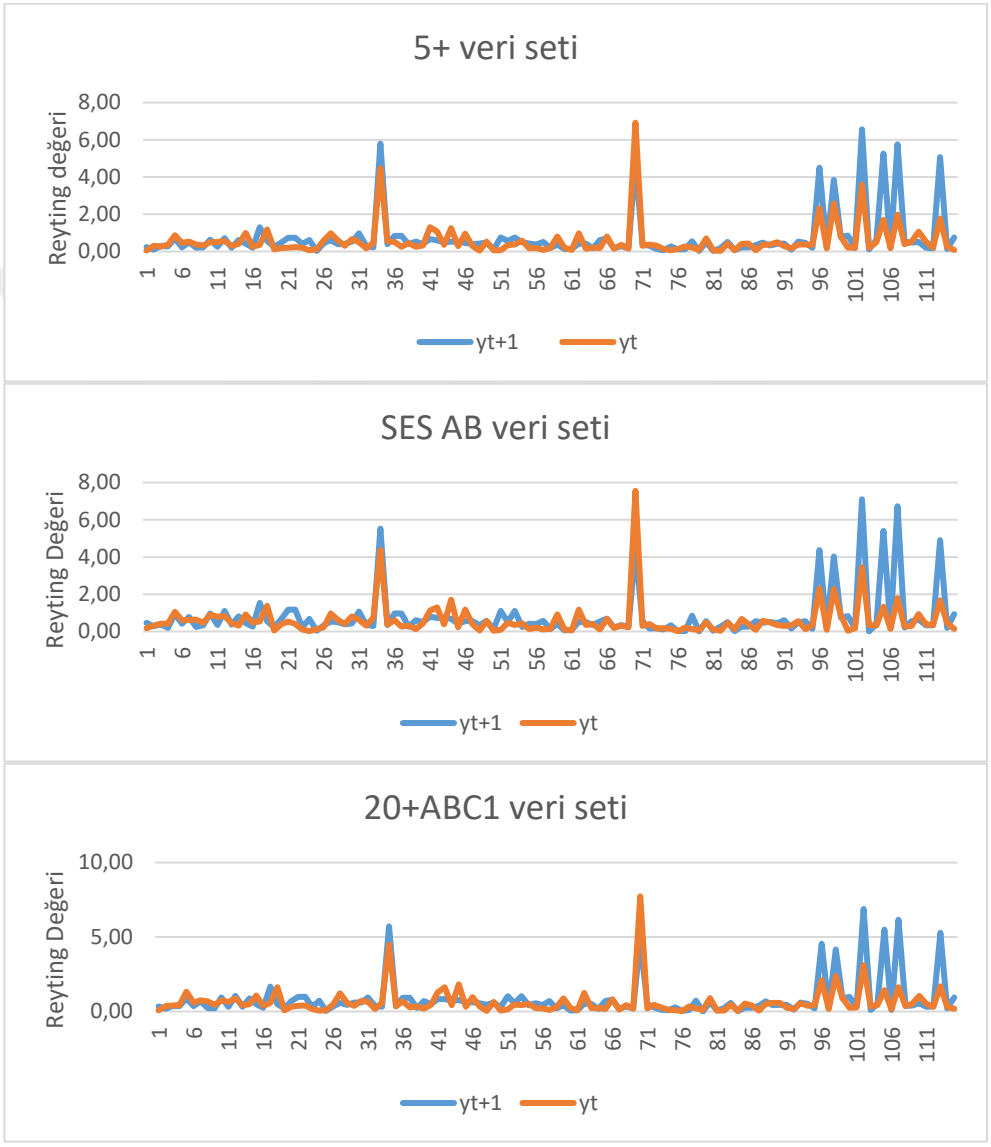
5.6. Bulgular

Sosyal medyadan elde edilen kullanıcı görüşlerinin gelecek tahmininde kullanılabilirliğini araştıran bu çalışma kapsamında işletme olarak seçilen televizyon kanalının ürünleri olan programlar hakkında Twitter'dan seyirci görüşleri toplanmıştır. Bu iletileri sayısal değerlere çevirerek gelecek tahmininde kullanılması sağlanmıştır. Sosyal medya verilerinin dışında programlara has özelliklerin de tahmin edici değişkenler olarak kullanıldığı tahmin modeli ile gelecek programın reyting değeri tahmin edilmeye çalışılmıştır. Oluşturulan program reyting veri setinin tanımlanması amacıyla yapılan incelemelerde gelecek bölüm reyting sınıfı a olan 570, b olan 1807, c olan 305, d olan 144 ve e olan 249 program olduğu gözlenmiştir. Kasım 2015- Haziran 2016 aralığındaki 8 ayı kapsayan yayın akışındaki sinema filmleri ve çizgi filmlerin çıkarıldığı veri setinde 39 farklı programa ait bilgi bulunmaktadır. Bu programlardan 19'u geçen sezondan devam eden program niteliğinde iken 20 tanesi bu sezon yayına başlamıştır. 39 farklı program için 8 aylık süreçte derlenen 3075 yayının ilgili gün ve saatteki reyting değeri ile gelecek bölümde aldığı reyting değeri karşılaştırıldığında 878 tanesinin gelecek bölümde reytingini arttırmış olduğu, 2197 tanesinin ise reytinginin bir sonraki bölümde azaldığı görülmüştür. Programların aldığı tweet sayıları ve bunların olumlu, olumsuz ya da nötr olma sayıları karşılaştırıldığında ise 115 programın aldığı negatif tweet sayısı pozitif ve

nötr tweetlerden fazladır. Bu 115 program incelendiğinde ise %71'inin bir sonraki bölümde reyting değerinin düştüğü anlaşılmıştır (Şekil 23). Sadece verileri incelenmesi ile de görüldüğü gibi programlar hakkında atılan tweetler ve bunların duyguları geleceğe yönelim hakkında bilgi içermektedir. Öte yandan geleneksel reyting ölçme sistemi ile düşük reytinge sahip olan programlardan bazıları yayın süresince 100'ün üzerinde tweet almıştır ve ilgili programların reyting değerlerinin gelecek bölümde arttığı gözlenmiştir. Bu da geleneksel reyting ölçme sisteminin özellikle gece kuşağında yayınlanan ya da spor karşılaşmaları gibi özel içerikli yayınlarda yanıltıcı olabileceği bu nedenle seyirci görüşlerinin dikkate alınması gerektiğini ortaya çıkarmıştır. Literatürde "data drilling" olarak geçen veri setinde çeşitli noktalara odaklanılıp verinin küçültüle küçültüle derinlemesine araştırılması anlamına gelen yöntem ile yapılan incelemeler sayesinde sosyal medya verilerinin işletmeye önemli katkılar sağlayabileceği ortaya çıkarılmıştır.

Çalışmada yapılan duygu analizi sayesinde metin verileri olumlu, olumsuz veya nötr olarak sınıflandırılmıştır. Bu sayede iletiyi yazan kişilerin görüşleri gruplanmıştır. Duygu analizi çalışmaları çoğunlukla mühendislik alanında uygulansa da işletme alanında da büyük ihtiyaç duyulmaktadır. Fakat metinlerin ön işleme ve sınıflandırılması aşamalarında kodlama bilgisine sahip olunması gerekliliği az sayıda sosyal bilimci tarafından kullanılmasına neden olmaktadır. Bu çalışmada önerilen sözlük temelli yaklaşım hem işletmeye veya konuya göre oluşturulan sözlüklerle daha yüksek performanslı sınıflandırma yapılabilmesini hem de duygu analizi yönteminin üst düzey kodlama bilgisine sahip olmayan pek çok araştırmacı tarafından kullanılabilmesini sağlamaktadır.

Şekil 23: Negatif tweet oranı yüksek olan programların ilgili gündeki ve gelecek gündeki reyting değerleri



Bunların yanı sıra çalışmada, uzman görüşü ile kategorize edilen reyting değerleri ile kümeleme analizi sonuçlarının birbirileri ile ilişkili olup olmadıkları test edilmiştir. Kümeleme analizinde 8 aylık yayın süresince eksik ve hatalı verilerin çıkarılması ile yaklaşık 309169 satırlık 36 sütunluk veri seti kullanılmıştır. Bu veri seti her dakika için ilgili program kaç tweet aldı bunların kaçı negatif, nötr ya da pozitif, favori edilenlerin sayısı gibi sadece sosyal medya verilerinden oluşmaktadır. İlgili veri setinin kümelere

ayrılmasındaki amaç hangi dakikaların birbiri ile benzer özelliklere sahip olduğunun belirlenmesidir. Geleneksel sistem ile ölçülen reyting değerlerinin uzman görüşü ile kategorize edilmesiyle de benzer amaç güdülmüştür. Dolayısı ile analitik olarak ayrışan kümeler ile uzman görüşüyle kategorize edilen sınıf değerlerinin ilişkili olup olmadığı araştırılmıştır. Böylece geleneksel yöntemle hesaplanan reyting sistemi ile Twitter'da kanal ve program hakkında yazılan iletilerin analizinin de ilişkili olup olmadığı ortaya çıkarılmaya çalışılmıştır. Elde edilen sonuçlar, geleneksel reyting değerlerinin kategorizasyonu ile kümeleme analizi ile 5 küme oluşturulmasının birbiriyle alakalı olduğunu göstermiştir.

Tahmin modelinde geçmiş çalışmalardaki performans başarısı nedeniyle seçilen üç veri madenciliği yöntemi – karar ağacı, destek vektör makineleri, yapay sinir ağları- uygulanmıştır. Bu sayede eğitim setleri algoritmalar vasıtasıyla eğitilmiş ve test setlerindeki veriler için çıktı değerleri tahmin edilmeye çalışılmıştır. Algoritmalar çıktı değişkenini etkileyen girdiler/öznitelikler hakkında bilgi verme özelliğinin dışında yapılan tahminin başarısını da vermektedir. Doğruluk oranı adı verilen tahmin performans ölçütü olarak sıklıkla kullanılan orana göre algoritmaların karşılaştırılması mümkün olmaktadır. Tablo 24-26 her veri seti için kullanılan tüm algoritmaların en uygun parametre değerleri ile sağladıkları doğruluk oranlarını göstermektedir.

Tablo 24: 5+ veri seti için en iyi sonuçlar tablosu

5+		
Yöntem	Fonksiyon/Parametre	Doğruluk Oranı
Karar ağacı	C5.0	0.8032
Destek vektör makineleri	Polinom/d=3/c=100/ γ=0.01	0.8081
Destek vektör makineleri	RTF/c=10/ γ=0.01	0.8260
Destek vektör makineleri	Lineer	0.8160
Yapay sinir ağları	Geriye Yayılım/n=10/l=0.01/mc=0.3	0.8390

Tablo 25: SES AB veri seti için en iyi sonuçlar tablosu

SES AB

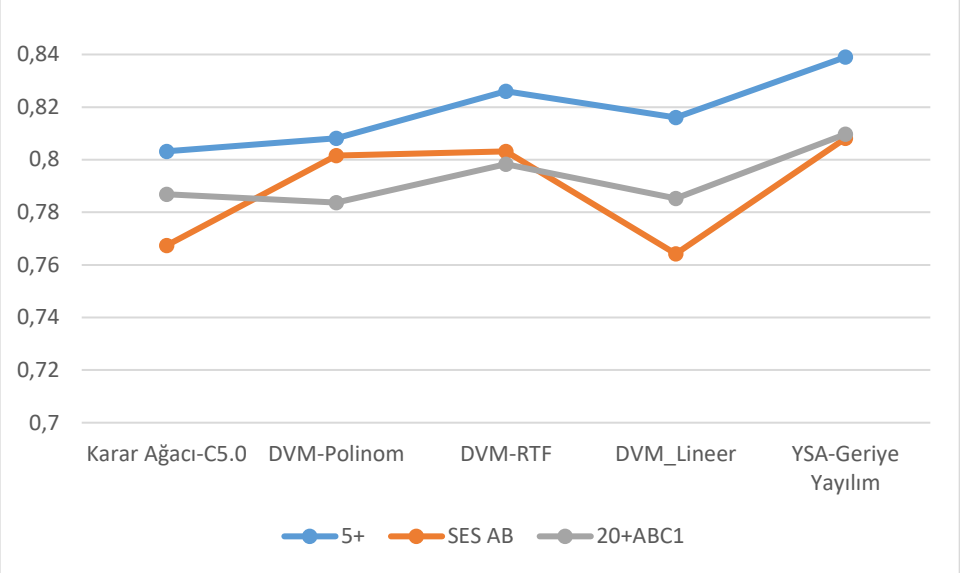
Yöntem	Fonksiyon/Parametre	Doğruluk Oranı
Karar ağacı	C5.0	0.7674
Destek vektör makineleri	Polinom/d=3/c=100/ γ=0.01	0.8016
Destek vektör makineleri	RTF/c=10/ γ=0.01	0.8032
Destek vektör makineleri	Lineer	0.7642
Yapay sinir ağları	Geriye Yayılım /n=10/l=0.01/mc=0.3	0.8081

Tablo 26: 20+ABC1 veri seti için en iyi sonuçlar tablosu

20+ABC1		
Yöntem	Fonksiyon/Parametre	Doğruluk Oranı
Karar ağacı	C5.0	0.7869
Destek vektör makineleri	Polinom/d=3/c=100/ γ=0.01	0.7837
Destek vektör makineleri	RTF/c=10/ γ=0.01	0.7983
Destek vektör makineleri	Lineer	0.7853
Yapay sinir ağları	Geriye Yayılım /n=10/l=0.01/mc=0.3	0.8097

5+ veri seti ile uygulanan tüm algoritmaların doğruluk oranları %80 ile %83 aralığında değişmektedir. En iyi performansı yapay sinir ağları algoritması gösterirken en düşük doğruluk oranı karar ağacı algoritması ile elde edilmiştir. SES AB veri setinde algoritmaların sağladıkları doğruluk oranları %76 ile %80 arasında değişiklik göstermektedir. Sırasıyla en başarılı tahmin performansına sahip algoritma yapay sinir ağları iken destek vektör makineleri/ lineer fonksiyon algoritması en düşün doğruluk oranını vermiştir. 20+ABC1 veri setinde ise doğruluk oranlarının %78 ile %80 arasında olduğu gözlenmiştir. Diğer veri setlerinde olduğu gibi en yüksek doğruluk oranına sahip algoritma yapay sinir ağlarıdır. Destek vektör makineleri/ polinom fonksiyon algoritması ise en düşük tahmin performansına sahiptir. Şekil 24'de üç veri seti için uygulanan tüm algoritmalara göre elde edilen doğruluk oranlarını özetlenmiştir.

Şekil 24: Algoritmalara göre doğruluk oranları



Tablo ve grafikte görüldüğü üzere üç veri setinde de en yüksek tahmin performansına sahip algoritma yapay sinir ağları olmuştur. İkinci en yüksek tahmin performansına sahip algoritma olan destek vektör makineleri/ radyal temelli fonksiyon da her üç veri setinde yapay sinir ağlarından sonra en yüksek doğruluk oranını sağlamıştır. Doğruluk oranları arasında farkların çok küçük olmasından dolayı kötü performansa sahip olan bir algoritmadan bahsetmek mümkün değildir. Uygulanan tüm algoritmalar anlamlı performans değerlerine sahiptir, dolayısı ile karar ağacı, destek vektör makineleri ve yapay sinir ağları algoritmalarının gelecek program reyting değeri tahmininde kullanılmasının uygun olduğu ortaya çıkarılmıştır. Öte yandan sosyal medya verilerinin nitelik ve nicelik olarak analiz edilerek işletmeler için değer yaratabilecek anlam ifade eden tahminler yapılmasının mümkün olduğu bulgusuna ulaşılmıştır. Çalışma kapsamında incelenen televizyon kanalı için geleneksel yöntemlerle ölçülen ve ana kütleye göre üç farklı gruba ayrılmış reyting değerlerinin her biri için de gelecek programın reyting sınıfının tahmin edilebileceği sonucuna ulaşılmıştır.

SONUÇ

Teknolojinin gündelik hayatın önemli bir parçası haline gelmesi bilinen anlamdaki veri kavramının değişmesine neden olmaktadır. Paylaşılan fotoğraflar, yazılan iletiler, konum bilgileri, ziyaret edilen web sayfaları gibi kişisel cihazlarla yapılan tüm aktiviteler, makine ve sensör verileri gibi çok çeşitli kaynaklardan çeşitli tiplerdeki veriler kayıt altında tutulmaktadır. Bu da veri boyutlarında büyümeye, yapısal olmayan ya da yarı yapısal formatta veri tiplerinin ortaya çıkmasına, verinin oluşma hızında artışa neden olmakta ve analiz edilerek değer yaratma imkânı sunmakta, yani büyük veri kavramını ortaya çıkarmaktadır. Kişisel cihazların yanında sanayi ve hizmet alanlarında da teknolojik altyapıların gelişmesi daha akıllı cihazların kullanılması ile büyük veri hacimlerine kolaylıkla ulaşılmaktadır. Makine verileri, sensör verileri vb. çeşitliliğe sahip büyük hacimli veri yığınlarının oluşması bunların anlamlandırılması ihtiyacını doğurmuştur. Dolayısı ile büyük verinin sadece verinin boyut, çeşitlilik ve diğer özellikleri ile tanımlanması eksik kalmaktadır. Büyük veri gelişmiş bilgisayar teknolojileri vasıtasıyla eldeki büyük hacimli, çeşitliliği yüksek, hızlı bir şekilde oluşan verilerin depolanıp analiz edildiği bir sitemdir. Bu sistem değersiz yığınlar olarak görülen verilerden anlam çıkarılarak değer yaratılmasını sağlamaktadır.

Öte yandan gelişen teknoloji geleneksel istatistikten veri analitiğine dönüşümün en önemli nedenlerinden birisidir. İstatistiksel yaklaşım araştırma sorusunun belirlenmesi, bu sorulara göre hipotezlerin geliştirilmesi, verinin toplanması ve kuramın geçerliliğinin araştırılması üzerine kurulduğundan hipotezden kurama ulaşmak hedeflenmektedir. Teknolojik gelişmeler ise bilimsel yaklaşımları, veri setlerine göre araştırma sorularının belirlendiği, ilgili veri setlerinin içerdiği ilişkilerin araştırıldığı, hipoteze dayalı yöntemlerden keşfetmeye dayalı yöntemlere doğru kaydırmaktadır. Geçmişte araştırma sorularına göre düzenlene deneyler ile veri toplanırken, günümüzde gelişen teknoloji ile dakikada binlerce verinin toplanması mümkün olmaktadır. Bu da veriden araştırma sorusu çıkarılmasına neden olmuş dolayısı ile istatistiksel yaklaşımların geliştirilmesine ya da farklı yaklaşımların ortaya çıkarılması ihtiyacını doğurmuştur. Bunun yanı sıra verinin

boyutu büyüdükçe hipotez testlerindeki her sorunun istatistiksel olarak anlamlı çıkmaya eğilimli olması da bu ihtiyacı arttırmıştır.. Dolayısıyla istatistiksel metotlar ve varsayımlarının artan veri boyutları ve çeşitliliği karşısında etkinliğini kaybetmeye başlaması istatistiksel modeller, metotlar, teoriler ve bilgisayar gücünü bir araya getiren çok disiplinli bir alan olan veri analitiğinin/biliminin ortaya çıkmasına yol açmıştır.

Teknolojik gelişmeler verinin boyut ve çeşitliliğini arttırsa da aynı zamanda daha güçlü bilgisayar ve çözüm teknolojileri de ortaya çıkmıştır. Bu da veri setinin bütününü kullanabilen, geçmişte yapılanlardan daha derin analizlerin yapılabildiği algoritmaların geliştirilmesini bu verilerin işlevsiz yığınlar olmadığının ortaya çıkarılmasını sağlamıştır. Böylece veri analitiği kavramının gerek çözüm teknikleri gerek veri toplama ve depolama teknikleri bakımından geliştirilmesi ile büyük veri analitiği kavramı ortaya çıkmıştır.

Veri analitiğinin ve gelişen adı ile büyük veri analitiği, işletmenin tüm fonksiyonlarında önemli faydalar yaratmaktadır. Bu nedenle bir alt dal ve disiplin olarak işletme analitiği kavramı ortaya çıkmıştır. İşletme analitiği genel tanımı itibariyle; işletmelerin içyapısında biriken veya işletmeler için önemli olan dış kaynaklardan toplanan verilerin derlenmesi, depolanması, düzenlenmesi ve analiz edilmesiyle anlamlı bilgiler ortaya çıkararak, işletmeler için fayda ve değer yaratılmasıdır. Diğer bir deyişle işletme analitiği, veri analitiği aracılığıyla vakaların anlaşılması için kullanılan prensipler, süreçler ve teknikler bütünüdür. Veri temelli kararlar; tecrübeler ve sezgiler yerine verinin analiz edilmesi ile alınan kararlardır. Bu sayede işletmeler müşteri bağlılığından, eleman seçimine, satış tahmininden, yatırım kararına çeşitli alanlarda verilerin analiz edilmesiyle başarılı kararlar alabilmektedirler.

İşletme analitiğinin sosyal medya analitiği ile birleştiği noktada, sosyal medyadan elde edilen veriler ile analiz yapılması, uygulama alanı ve yaratacağı faydalar ve değerler üzerinde yoğunlaşılmıştır. Sosyal medya kanallarından, anket ya da diğer görüşme türleri ile elde edilemeyecek sayıda kişinin görüşünü çok kısa sürelerde elde etmek mümkündür. Fakat sosyal medya verileri bilinen istatistiksel yöntemlerde kullanılan veri tiplerinden

farklıdır. Yarı yapısal formattaki bu veri seti içeriğinde derin anlamlar saklayabilmektedir. Geleneksel istatistiksel yöntemlerde veri elde etmeden önce araştırma sorusu belirlenir, araştırmacı sorularını önceden belirler ve bu yapının dışına çıkılmaz. Sosyal medyadan elde edilen veriler ise kendi araştırma sorularını kendi ortaya çıkarabilmektedir. Dolayısı ile fayda ve değer yaratmak için veri setini iyi analiz edilmesi gerekmektedir. Metin sınıflandırma, duygu analizi, kelime bulutları veya birliktelik analizi gibi yöntemler sonuçları ile sağladıkları değerin yanı sıra veriyi yapısal hale dönüştürüp başka analitik çalışmalarında girdi olarak kullanılmasını da sağlamaktadır. Bu sayede müşterilerin yazdıkları iletiler işletme için daha kolay anlaşılır ve kullanılabilir hale gelebilmektedir. Bu da işletmenin yapacağı yatırımların, pazarlama stratejilerinin, geliştireceği ürünlerin belirlenmesi gibi aşamalarda karar desteği sunmaktadır. Küresel çapta büyük şirketler çalışanlarının performanslarını artık yıllık ya da belli periyotlarla değil, anlık olarak değerlendirmeye başlamışlardır. Bir başka deyişle, paradigmalar evrilmektedir; periyodik ve kesikli yaklaşımlar yerini, anlık ve sürekli bakış açılarına bırakmaktadır. Sosyal medya analizleri de, tüketici davranışlarının öngörülmesi bağlamında post modern bir yaklaşımla, "süreklilik" nosyonuna hizmet etmektedir. Müşterilerin sosyal medyada beğeni ile bahsettikleri ürünlerin piyasadan kaldırılması, belirli bir ürün hakkında sosyal medyadan yapılan şikâyetlerin göz ardı edilmesi, ürün ile ilgili geliştirilmesi/eklenmesi gereken özellikleri ileten müşteri görüşlerinin kaçırılması gibi sosyal medya verilerinin analiz edilmemesinden kaynaklı hatalar gün yüzüne çıkabilmektedir. Böylece müşteri geribildirimleri tüm boyutlarıyla daha çok anlaşılmaya çalışılacak, müşteriler 360 derece yaklaşımıyla daha iyi analiz edilecek ve müşteri davranışının öngörülmesi daha nitelikli bir şekilde sağlanabilecektir. Bu amaçla metin verilerine duygu analizi teknikleri uygulayarak yazan kişinin ruh hali, düşüncesi ya da fikri ortaya çıkarılabilmektedir. Literatürde yer alan yoğun programlama bilgisi kullanılan duygu analizi çalışmaları yüksek performanslar ile sonuçlar üretmektedir ve birçok mühendislik temelli araştırmacı bu performansları arttırmaya yönelik algoritmalar üzerine çalışmaktadır. Fakat ilgili çalışmalar uygulama zorluğu nedeniyle her alandan araştırmacıların kullanabileceği yapıya sahip değillerdir. Bu da metinlerde gizli kalmış duygularının açığa çıkarılması ile

yaratılacak değerlerden yoksun kalınmasına neden olmaktadır. Bu çalışma kapsamında geliştirilen algoritma ise uygulama kolaylığı ön planda olan performans kalitesi yüksek bir duygu analizi algoritması önermektedir. Algoritmada kullanım ve erişim kolaylığı özellikleri nedeniyle, sözlük oluşturma aşamasında R programlama dili, sözlük skoru belirleme ve karşılaştırma aşamalarında ise Excel ve VBA kullanılmıştır. Sözlük temelli duygu analizi algoritmalarında kritik konulardan biri sözlük oluşturma aşamasıdır. Her veri seti kendine has jargonlara sahip olabilmektedir görüşünden yola çıkarak bu çalışmada yarı denetimli bir teknikle sözlük oluşturulmuştur. Bu amaçla 1000 adet tweet pozitif olarak 1000 adet tweet de negatif olarak etiketlenmiştir. Sonrasında R programlama dili aracılığı ile veri seti metin ön işleme sürecine tabii tutulmuş; negatif ve pozitif etiketli veri setlerinde ayrı ayrı en sık geçen 1000 kelime belirlenmiş ve bunlardan anlamlı olanları ayıklanarak pozitif duygu sözlüğü ve negatif duygu sözlüğü oluşturulmuştur. Ardından veri setindeki tüm iletilerin pozitif duygu sözlüğünden ve negatif duygu sözlüğünden kaçar kelime içerdikleri belirlenmiş ve sayıca yüksek olan duyguya etiketlenmiştir. Önerilen duygu analizi algoritması ile işletme ve ürünler hakkında sosyal medyadan toplanan verilerin, iletiyi yazan kişilerin ilgili ürünler ve işletme hakkındaki görüşlerinin olumlu, olumsuz ya da nötr oldukları araştırılmıştır

Yeni piyasaya sürülen bir ürünün ilk zamanlar satış rakamları düzensiz olması karar vermeyi zorlaştırırken, sosyal medyadan kullanıcı görüşlerinin takip edilmesi ile ilgili ürün hakkında daha hızlı kararlar verilebilir. Ya da uzun yıllardır piyasada bulunan bir ürünün satış değerlerinin dalgalı ya da düşüşte olmasının nedenleri gene sosyal medyadan elde edilen kullanıcı görüşleri ile analiz edilebilir ve işletmeye ürün hakkında strateji geliştirmeye yardımcı olunabilir. Çalışmada da bu bulguların elde edilebilmesi hedeflenmiştir. Duygu analizi ile elde edilen bilgi, ileti sayısı ve ürünlerin kendine has diğer özellikleri ile birlikte tahmin modelinde girdi olarak kullanılması ile işletmeye ilgili ürünler hakkında karar alınırken faydalanabilecekleri bir sistem önerilmiştir. İşletme olarak bir televizyon kanalı seçilmiştir. Kanalda yayınlanan programlar, ürünleri/hizmetleri temsil etmektedir. Dolayısı ile televizyon kanalındaki kullanıcı/müşteri konumundaki seyircilerin programlar hakkındaki görüşleri derlenmeye

çalışılmıştır. Bu sayede, satış rakamlarını temsil eden geleneksel yollarla ölçülen reytinglerin tahmin edilmesi hedeflenmiştir.

Müşterilerin sesi olarak adlandırılan kavramın yani kullanıcıların ürün ya da hizmetler hakkındaki görüşleri, istekleri, geri bildirimleri günümüzde sosyal medyadan elde edilebilmektedir. Televizyon kanalları için de yayınlanan programların başarısı için en önemli ölçüt geçmişte geleneksel yollarla ölçülen reyting değerleriydi. Şimdi ise sosyal medya sayesinde kanal yöneticileri seyircilerinin sesini daha rahat duyabilmekte ve direkt olarak seyirciden programlar hakkındaki görüşlerini öğrenebilmektedir. Sosyal medya kanallarının yaygınlaşması, teknolojinin daha fazla kullanıcının ulaşabileceği hale gelmesi, sosyal medyada biriken verilerin kullanılabilirliği üzerine çalışmaları geliştirmiştir. Bu çalışmada da benzer bir araştırma sorusu üzerinden yola çıkılmıştır. Geleneksel televizyon reyting ölçümlerinin, sosyal medyadan elde edilen seyirci görüşleri ile tahminlenip tahminlenemeyeceği araştırılmıştır. Bu amaçla Twitter'dan yarı yapısal formatta olan büyük veri setleri elde edilmiştir. Bu sayede seyircilerin yayın akışındaki programlar hakkındaki görüşleri toplamış ve analiz edilerek bu görüşlerin olumlu olumsuz ya da nötr olduğunu araştırılmıştır. Duygu analizi kapsamında geliştirilen duygu durumu belirleme algoritması sayesinde yaklaşık 1.200.000 tweet metninin hangi duyguyu barındırdığı belirlenebilmiştir. Türkçe dili için geliştirilmiş duygu analizi aracı olmadığından ve üzerinde çalışılan algoritmaların kolay uygulanabilir olmamasından dolayı geliştirilen sözlük temelli algoritma önerilmiştir. Önerilen algoritma sayesinde her alandan araştırmacı çalıştığı veri setine uygun duygu sözlüğünü oluşturup yoğun kodlama bilgisine ihtiyaç duymadan kolaylıkla duygu analizi çalışması yapabilmektedir. Yapılan duygu analizi çalışması ile pozitif, negatif veya nötr olarak etiketlenen metin verileri yapısal hale getirilmiştir. Sosyal medyada kanal ve programları hakkında yazılan iletilerin sayısı ve bu iletilerin duygularının gerçek reyting değerleri ile ilişkili olup olmadığının incelenebilmesi için bu verilere kümeleme analizi uygulanmıştır. Kümeleme analizi ile dakikalık olarak atılan tweet sayıları bunların pozitif, negatif veya nötr duygulu olanlarının sayıları, retweet ve favori edilme sayıları gibi özniteliklerin gruplanması sağlanmıştır. Eş zamanlı olarak kanal yöneticilerinden reyting sınıfları hakkında bilgi

alınmış ve reyting değerlerinin ilgili kanalda 5 farklı kategoriye ayrıldığı öğrenilmiştir. Uzmanlarca kategorize edilen ve kümelenen bu iki grubun birbiri ile ilişkisinin incelenmesi için geliştirilen hipotez ki-kare bağımsızlık testi ile araştırılmış ve küme sonuçları ile gerçek reyting sınıfları arasında bir ilişki olduğu, bunların birbirinden bağımsız değişkenler olmadığı sonucuna ulaşılmıştır.

Çalışmada duygu analizi sonuçlarının ve programlara has özelliklerin girdi olarak kullanıldığı tahmin modeli ile geleneksel reyting ölçümlerine göre belirlenen reyting değerlerinin gelecek program için tahmin edilmesi amaçlanmıştır. Bu amaçla, karar ağacı, destek vektör makineleri ve yapay sinir ağları veri madenciliği yöntemleri olarak seçilmiştir. Her üç yöntem de R programlama dili aracılığı ile uygulandığında tahmin performansları bakımından hepsinin anlamlı olduğu anlaşılmıştır. Böylece sosyal medya verilerinin nicelik ve nitelik (duygu analizi aracılığıyla) değerleri ile ürünlere has özelliklerin gelecek dönem reyting tahmininde kullanılabilir olduğu ve sonuçların yorumlanabileceği ortaya çıkarılmıştır. Tahmin modeli, 5+, SES AB ve 20+ABC1 olmak üzere üç farklı veri setinde uygulanmıştır. Bunun nedeni, geleneksel reyting ölçüm sisteminin sosyoekonomik statü gruplarına göre ölçüm sonuçları vermesi ve çalışmada bahsi geçen kanalın bu üç gruba göre reytingleri incelemesidir. Model üç veri seti içinde karar ağacı, destek vektör makineleri ve yapay sinir ağları ile analiz edilmiş ve sonuçlar aşağıda paylaşılmıştır.

5+ veri seti, 5 yaş üzeri tüm bireylere göre hesaplanan reyting değerlerini sınıf değeri olarak kullanmıştır. Karar ağacı algoritması ilgili veri setinde yapılan tahminde %80.32'lik doğruluk oranı sağlamıştır. Ayrıca program adı ile pozitif duygulu atılan tweetler, program türü, hitap ettiği kitle bilgisi, program süresince atılan toplam tweet sayısı ve hangi gün yayınlandığı bilgilerinin sınıf değerinin belirlenmesinde etkili olduğu ortaya çıkarılmıştır. Bu sayede kanal yöneticilerinin alacağı kararlarda sosyal medyadan elde edilen verilerin kullanılmasının değer yaratacağı anlaşılmaktadır. Destek vektör makineleri ilgili veri setinde en iyi performansı %82.60 ile radyal temelli fonksiyonun kernel fonksiyon olarak seçildiği algoritmayla vermiştir. Yapay sinir ağları algoritması ise

%83.9 doğruluk oranı ile ilgili veri seti için en yüksek doğruluk oranına sahip algoritma olmuştur. Bu algoritmaya göre ise sınıf değerinin belirlenmesinde önem arz eden girdilerden bazıları, program türü, programdaki ünlülük seviyesi, program adı ile pozitif duygulu atılan tweetler, hangi gün yayınlandığı ve program başlamadan 2 saat önce ve bittikten 2 saat sonra aralığında atılan toplam tweet sayısıdır.

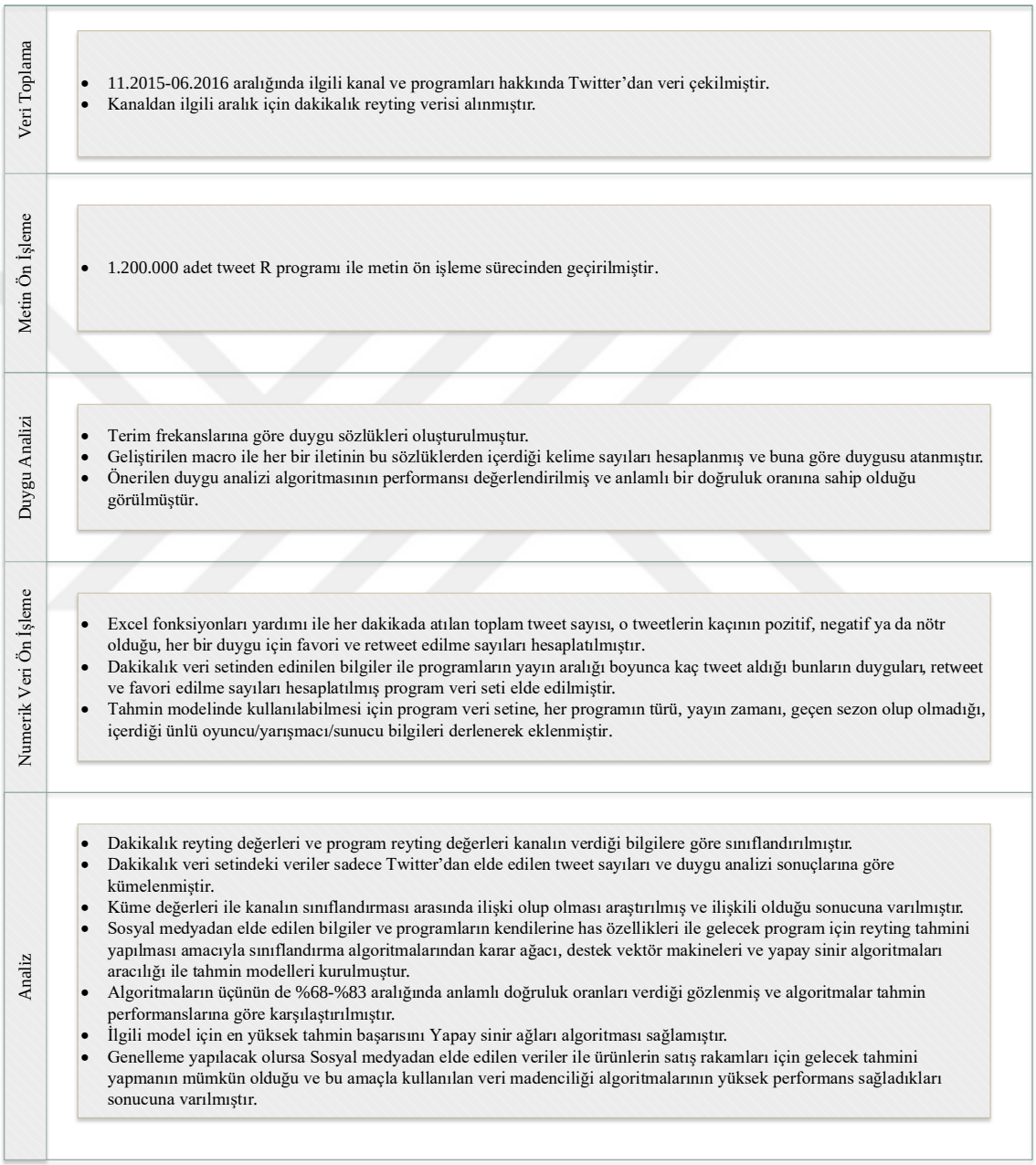
SES AB veri seti, sosyoekonomik statü gruplarından A ve B'de yer alan bireylere göre hesaplanan reyting değerlerini sınıf değeri olarak kullanmıştır. Karar ağacı algoritmasının %76.74 doğruluk oranını sağladığı analiz, sınıf değerinin belirlenmesinde programın hitap ettiği kitle, programın türü, program başlamadan 2 saat önceden ve bittikten 2 saat sonraya kadar olan aralıkta atılan toplam tweet sayısı, kanal adı ile atılan pozitif duygulu tweet sayısının önemli olduğunu ortaya çıkarmıştır. Destek vektör makineleri algoritmasında kernel fonksiyon olarak radyal temelli fonksiyon kullanıldığında %80.32 doğruluk oranına ulaşılmıştır. Bu oran uygulanan tüm kernel fonksiyonları içerisinde en yüksek doğruluk oranıdır. Yapay sinir ağları algoritması ise tüm algoritmalar içerisinde en yüksek doğruluk oranı olan %80.81'i sağlamıştır. Bununla beraber algoritma program türünün, yayınlandığı günün, geçen sezon var olup olmamasının, programdaki ünlülük seviyesinin, yayın aralığının ve program adı ile atılan toplam tweet sayısının sınıf değerinin belirlenmesinde önemli olduğunu ortaya çıkarmıştır.

20+ABC1 veri seti ise, 20 yaş üzeri ve sosyoekonomik statü gruplarından A, B ve C1'de yer alan bireylere göre hesaplanan reyting değerlerini sınıf değeri olarak kullanmıştır. Tahmin modeli için ilk olarak karar ağacı algoritması kullanılmış ve %78.69'luk doğruluk oranına ulaşılmıştır. Algoritma sınıf değeri belirlenirken önem arz eden değişkenleri kanal adı ile atılan pozitif duygulu tweetlerin retweet sayısı, programın yayınlandığı gün, kanal adı ile atılan toplam tweet sayısı, kanal adı ile atılan pozitif duygulu tweet sayısı ve program başlamadan 2 saat önceden ve bittikten 2 saat sonraya kadar olan aralıkta atılan toplam tweet sayısı olarak belirtmiştir. Destek vektör makineleri algoritması ile model uygulandığında diğer veri setlerindeki gibi radyal temelli fonksiyon diğer fonksiyonlara göre daha yüksek performans göstermiştir. Böylece algoritma %79.84 olasılık ile yeni

gelecek bir verinin sınıf değerini doğru tahmin edebilmektedir. Yapay sinir ağları algoritması ise %80.97 ile en yüksek tahmin performansına sahiptir. Algoritmaya göre ilgili veri setinde sınıf değerinin belirlenmesinde program türü, geçen sezon olup olmaması, program süresince kanal adı ile atılan Pozitif duygulu tweetlerin favori edilme sayısı ve program başlamadan 2 saat önceden ve bittikten 2 saat sonraya kadar olan aralıkta atılan toplam tweet sayısı önem arz etmektedir.

Aşama aşama sonuçları Şekil 25'de özetlenen tez çalışması kanala test yayınında olan ya da devam edip etmeyeceği konusunda düşünülen programlar hakkında alınacak kararlara destek mekanizması oluşturmayı hedeflemiştir ve yapılan analizler sonucunda bu hedefe ulaşıldığı görülmüştür. Mevcut yapıda reyting değerlerine bakılarak programların iyi gittiği ya da düşüşte olduğu kararı verilirken, bu çalışma ile reyting ölçüm sisteminin tüm evreni temsil gücünün zayıf olabileceği bu yüzden sosyal medyadan izleyici görüşlerinin elde edilmesinin daha etkili olacağı üzerinde durulmuştur. İzleyicilerin görüşleri ve programların kendilerine has özelliklerinin birlikte incelendiği bu çalışma, kanala yayından kaldırmayı planladığı programlar ve yayınlaması gereken programlar hakkında bilgi sağlamaktadır. Bu bilgi programların başarısını sadece geleneksel yöntemlerle ölçülen reyting değerleri ile değerlendirmenin kanalı yanlış kararlar almaya sürükleyebilecek olmasından dolayı önemlidir. Bir programın gündüz kuşağında yayınlanması ile akşam kuşağında yayınlanması reyting değerlerinde önemli bir farka yol açmaktadır. Bunun yanı sıra sadece kadınlara yönelik olması veya 5 yaş üzeri tüm bireylere yönelik olması, programda oldukça ünlü olan kişilerin yer alması veya daha az tanınan kişilerin yer alması ve hafta içi yayınlanması veya hafta sonu yayınlanması reyting değerlerinde büyük farklara yol açmaktadır. Bu nedenle sadece reyting değerine göre karar almak yerine programa has özelliklerin ve seyirci görüşlerinin incelendiği bu çalışma kanala değer sağlamaktadır. Örnek vermek gerekirse kanal kurulan model sayesinde bugün yayınlanan a programının yarınki reyting değerini öngörebilmekte bu sayede programın gidişatına müdahale edebilmektedir. Öte yandan yüksek reyting alan programların sahip olduğu ortak özellikler ortaya çıkarılarak yeni yayınlanacak programların bu özelliklere sahip olacak şekilde düzenlenmesi mümkün olacaktır.

Şekil 25: Genel akış-sonuçlar



Test yayını aşamasında olan yeni bir programa devam edip etmeme kararını sadece reyting değerlerine bakarak değil izleyici görüşleri ve diğer özelliklere göre de karar verebileceklerdir. Bunların yanı sıra çalışmada tahminlenen reytinglere göre kanal reklam fiyatlarını düzenleyebilir. Reyting değerlerinin yüksek olacağının tahminlendiği

programlarda yayınlanacak reklamlardan farklı satış fiyatları talep edebilme imkânı ortaya çıkabilecektir. Veri seti ve tahmin modeli sonuçları incelendiğinde 2015-2016 sezonunda kanalın karşılaştığı ve çalışmada anlatılan sistem kullanılmış olsaydı daha hızlı karar alınabilecek olan bazı vakalar olduğu gözlenmiştir. Bunlardan biri olan X programı ilgili sezonda yeni başlamış, gündüz kuşağında, kadınlara yönelik bir programdır ve reyting değerleri benzerlerine göre önemli farklılığa sahip değildir. Fakat yayın süresince atılan tweetler incelendiğinde olumsuz tweetlerin daha fazla olduğu görülmüştür. Bu tweetler detaylı incelendiğinde seyircilerin, programın kadınları aşağıladığını ve bu nedenle programın iyi olmadığını belirttikleri anlaşılmıştır. Aynı zamanda tahmin modeli de program için gelecekte düşük reyting alacağı tahminini yapılmıştır. Fakat kanal seyircilerden gelen tepkileri anlık olarak takip edememiş ve program ancak bir buçuk ay sonunda yayından kaldırılmıştır.

Çalışmanın genel sonucuna bakılacak olursa; müşterisi sosyal medya kullanıcısı olabilecek herhangi bir işletme, ürünleri hakkında sosyal medyadan veri toplayıp bunlardan müşterilerinin görüşlerini öğrenebilir. Bu görüşleri doğrudan ya da geliştirilecek tahmin modellerinde dolaylı olarak kullanarak pazarlama, finans, üretim ve daha birçok alanda yeni stratejiler geliştirebilir ve işletmeye değer yaratabilir.

Gelecek çalışmalar;

- Geliştirilen tahmin modeli çeşitli sektörlerdeki birçok işletme için uygulanabilir.
- Tahmin modeline rakip işletmelerin bilgileri de eklenebilir ve rakip işletmeler için sonuçlar karşılaştırılabilir.
- Duygu analizi algoritmasının tahmin performansının arttırılacağı iyileştirmeler -basitliğinden ödün vermeden- yapılabilir.
- TV programları için geleneksel yöntemlerle ölçülen reyting sistemine sosyal medya verilerinin de eklendiği yeni bir ölçüm modeli geliştirilebilir.
- Geliştirilen tahmin modeli farklı veri madenciliği teknikleri ile denenebilir ve performansları karşılaştırılabilir.

KAYNAKÇA

AGRAWAL, Divyakant; et. al.: Challenges and Opportunities with big data 2011-1. 2011

AHKTER, Julie Kane; SORIA, Steven.: Sentiment analysis: Facebook status messages. Unpublished master's thesis, Stanford, CA, 2010.

AKARSU, Cenk; DIRI, Banu.: Turkish TV rating prediction with Twitter. In: Signal Processing and Communication Application Conference (SIU), 2016 24th. IEEE, 2016. p. 345-348.

AKPINAR, Haldun.: Data, Veri madenciliği- Veri Analizi, 2014.

ANDERSON, Chris.: The end of theory: The data deluge makes the scientific method obsolete. Wired magazine, 2008, 16.7: 16-07.

ARAMAKI, Eiji; MASKAWA, Sachiko; MORITA, Mizuki.: Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011. p. 1568-1576.

ARDIÇ, Burcu; GÖKTÜRK, Mehmet.: Kullanılabilir Uygulama Programlama Arayüzleri, 4. Ulusal Yazılım Mühendisliği Sempozyumu, Beşiktaş, İstanbul, 2009, 91-97.

ASUR, Sitaram; HUBERMAN, Bernardo A.: Predicting the future with social media. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. IEEE, 2010. p. 492-499.

BALOĞLU, Arzu,: Sosyal Medya Madenciliği, Beta Yayınları, 2015.

BATRINCA, Bogdan; TRELEAVEN, Philip C.: Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY, 2015, 30.1: 89-116.

BEKMAMEDOVA, Nargiza; SHANKS, Graeme.: Social media analytics and business value: A theoretical framework and case study. In: System Sciences (HICSS), 2014 47th Hawaii International Conference on. IEEE, 2014. p. 3728-3737.

BELLO-ORGAZ, Gema; JUNG, Jason J.; CAMACHO, David.: Social big data: Recent achievements and new challenges. Information Fusion, 2016, 28: 45-59.

BERGMEIR, Christoph; BENÍTEZ, José M.: Package 'RSNNS'. 2016.

BLISS, Catherine A., et al.: Twitter reciprocal reply networks exhibit assortativity with respect to happiness. Journal of Computational Science, 2012, 3.5: 388-397.

BODNAR, Todd; SALATHÉ, Marcel.: Validating models for disease detection using twitter. In: Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013. p. 699-702.

BOLLEN, Johan; MAO, Huina; ZENG, Xiaojun.: Twitter mood predicts the stock market. Journal of computational science, 2011, 2.1: 1-8.

BOSER, Bernhard E.; GUYON, Isabelle M.; VAPNIK, Vladimir N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. ACM, 1992. p. 144-152.

BOUKTIF, Salah; AWAD, Mamoun Adel.: Ant colony based approach to predict stock market movement from mood collected on Twitter. In: Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013. p. 837-845.

BOYD, Danah; CRAWFORD, Kate.: Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 2012, 15.5: 662-679.

BRAMER, Max.: Principles of data mining. London.: Springer, 2007.

BUDAK, D.; EL ABBADI, Amr.: Information diffusion in social networks: Observing and influencing societal interests. PVLDB, 2011, 4.12: 1-5.

CATAL, Cagatay; NANGIR, Mehmet.: A sentiment classification model based on multiple classifiers. Applied Soft Computing, 2017, 50: 135-141.

CHEN, Hsinchun; CHIANG, Roger HL; STOREY, Veda C.: Business intelligence and analytics: From big data to big impact. MIS quarterly, 2012, 36.4: 1165-1188.

CHENG, Mei-Hua; WU, Yi-Chen; CHEN, Ming-Chih.: Television Meets Facebook: The Correlation between TV Ratings and Social Media. American Journal of Industrial and Business Management, 2016, 6.03: 282.

CIOFFI-REVILLA, Claudio.: Computational Social Science (November 12, 2010). WILEY Interdisciplinary Reviews: Computational Statistics, Vol. 2, No. 3, pp. 259-271, May/June 2010. Available at SSRN: https://ssrn.com/abstract=1708051 or http://dx.doi.org/10.2139/ssrn.1708051

CLEVELAND, William S.: Data science: an action plan for expanding the technical areas of the field of statistics. International statistical review, 2001, 69.1: 21-26.

CODAL, Keziban SEÇKİN; COŞKUN, Erman.: Sosyal Ağ Türlerinin Karşılaştırılmasına İlişkin Bir Ağ Analizi. Abant İzzet Baysal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2016.

COREA, Francesco.: Can Twitter Proxy the Investors' Sentiment? The Case for the Technology Sector. Big Data Research, 2016, 4: 70-74.

CORTES, Corinna; VAPNIK, Vladimir.: Machine learning. Kluwer Academic Publishers, Boston, 1995, 20: 273-297.

CUKIER, Kenneth.: Data, data everywhere: A special report on managing information. Economist Newspaper, 2010.

CULOTTA, Aron.: Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the first workshop on social media analytics. ACM, 2010. p. 115-122.

CVIJIKJ, Irena Pletikosa; MICHAHELLES, Florian.: Online engagement factors on Facebook brand pages. Social Network Analysis and Mining, 2013, 3.4: 843-861.

ÇOBAN, Önder; ÖZYER, Barış; ÖZYER, Gülşah Tümüklü.: Sentiment analysis for Turkish Twitter feeds. In: Signal Processing and Communications Applications Conference (SIU), 2015 23th. IEEE, 2015. p. 2388-2391.

DA SILVA, Nadia FF; HRUSCHKA, Eduardo R.; HRUSCHKA, Estevam R.: Tweet sentiment analysis with classifier ensembles. Decision Support Systems, 2014, 66: 170-179.

DAVENPORT, Thomas H.: The human side of Big Data and high-performance analytics. International Institute for Analytics, 2012, 1-13.

DAVENPORT, Thomas H.; PATIL, D. J: Data Scientist: The Sexiest Job of the 21st Century-A new breed of professional holds the key to capitalizing on big data opportunities. But these specialists aren't easy to find—And the competition for them is fierce. Harvard Business Review, 2012, 70.

DEAN, Jared.: Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. John Wiley & Sons, 2014.

DIEBOLD, Francis X.: ’Big Data’Dynamic factor models for macroeconomic measurement and forecasting. In: Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society,”(edited by M. Dewatripont, LP Hansen and S. Turnovsky). 2003. p. 115-122.

DILLON, William R.; Multivariate analysis: Methods and applications. 1984.

GOLDSTEIN, Matthew.:

DING, Chao, et al.: The power of the “like” button: The impact of social media on box office. Decision Support Systems, 2017, 94: 77-84.

DODDS, Peter Sheridan, et al.: Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. PloS one, 2011, 6.12: e26752.

DRAGLAND, Asa.: Big data–for better or worse. SINTEF. no. 22 May 2013. Web. 27 Oct, 2013.

DU, Danyang; LI, Aihua; ZHANG, Lingling.: Survey on the applications of big data in Chinese real estate enterprise. Procedia Computer Science, 2014, 30: 24-33.

DUNHAM, Margaret H.: Data mining: Introductory and advanced topics. Pearson Education India, 2006.

EIRINAKI, Magdalini; PISAL, Shamita; SINGH, Japinder.: Feature-based opinion mining and ranking. Journal of Computer and System Sciences, 2012, 78.4: 1175-1184.

ELIAÇIK, Alpaslan Burak; ERDOGAN, Nadia.: Mikro Bloglardaki Finans Toplulukları için Kullanıcı Ağırlıklandırılmış Duygu Analizi Yöntemi. In: UYMS. 2015.

ELRAGAL, Ahmed.: ERP and Big Data: The Inept Couple. Procedia Technology, 2014, 16: 242-249.

EMANI, Cheikh Kacfah; CULLOT, Nadine; NICOLLE, Christophe.: Understandable big data: A survey. Computer science review, 2015, 17: 70-81.

EMIR, Senol; DINÇER, Hasan; A stock selection model based on fundamental and technical analysis variables by using artificial neural networks and support

TIMOR, Mehpare.: vector machines. Review of Economics & Finance, 2012, 2: 106-122.

EREVELLES, Sunil; FUKAWA, Nobuyuki; SWAYNE, Linda.: Big Data consumer analytics and the transformation of marketing. Journal of Business Research, 2016, 69.2: 897-904.

EVANS, James R.; LINDNER, Carl H.: Business analytics: the next frontier for decision sciences. Decision Line, 2012, 43.2: 4-6.

EYSENBACH, Gunther.: Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. Journal of medical Internet research, 2011, 13.4: e123.

FAN, Weiguo; GORDON, Michael D. T.: he power of social media analytics. Communications of the ACM, 2014, 57.6: 74-81.

FISHER, Bill; MILLER, Hayley.: Social Media Analytics. 2011.

FISHER, Danyel, et al.: Interactions with big data analytics. interactions, 2012, 19.3: 50-59.

GANDOMI, Amir; HAIDER, Murtaza.: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 2015, 35.2: 137-144.

GANTZ, John; REINSEL, David.: The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, 2012, 2007.2012: 1-16.

GHIASSI, M.; LIO, David; MOON, Brian. Pre-production forecasting of movie revenues with a dynamic artificial neural network. Expert Systems with Applications, 2015, 42.6: 3176-3193.

GHIASSI, Manoochehr; SKINNER, James; Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with applications, 2013, 40.16: 6266-6282.

ZIMBRA, David.:

GORDON, Jonathan; PERREY, Jesko; SPILLECKE, Dennis.: Big data, analytics and the future of marketing and sales. Forbes, New York, 2013, 22.

GRUHL, Daniel, et al.: The predictive power of online chatter. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005. p. 78-87.

GUNAWARDENA, Nipun, et al.: Instagram hashtag sentiment analysis. In: University of Utah CS530/CS630 Conference of Machine Learning. 2013.

GÜRSAKAL, Necmi.: Büyük Veri. 2014.

GÜRSOY, U. Tugba; BILGIN, Safiye.: BANKA MÜSTERILERININ INTERNET BANKACILIGINA ILISKIN YAKLASIMLARININ VERI MADENCILIGI TEKNIKLERI ILE INCELENMESI. Kafkas University. Faculty of Economics and Administrative Sciences. Journal, 2016, 7.14: 421.

HAIR, Joseph F., et al.: Multivariate data analysis . Uppersaddle River. Multivariate Data Analysis (5th ed) Upper Saddle River, 1998.

HANNEMAN, Robert A.; RIDDLE, Mark.: Introduction to social network methods. 2005.

HANSEN, Derek; SHNEIDERMAN, Ben; SMITH, Marc A.: Analyzing social media networks with NodeXL: Insights from a connected world. Morgan Kaufmann, 2010.

HASHEM, Ibrahim Abaker Targio, et al.: The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 2015, 47: 98-115.

HE, Wu; ZHA, Shenghua; LI, Ling.: Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 2013, 33.3: 464-472.

HEBB, Donald Olding.: The organization of behavior: A neuropsychological theory. Psychology Press, 2005.

HMEIDI, Ismail, et al.: Automatic Arabic text categorization: A comprehensive comparative study. Journal of Information Science, 2015, 41.1: 114-124.

HSU, Chih-Wei, et al.: A practical guide to support vector classification. 2003.

HUR, Minhoe; KANG, Pilsung; CHO, Sungzoon.: Box-office forecasting based on sentiments of movie reviews and Independent subspace method. Information Sciences, 2016, 372: 608-624.

IBM: Premier Healthcare Alliance IBM case study: IBM, 2012a

IBM: What is Big Data?, 2012b. Available from: ⟨ http://www-01.ibm.com/software /data/bigdata/what-is-big-data.html ⟩

ISSON, Jean-Paul; HARRIOTT, Jesse.: Win with advanced business analytics: creating business value from your data. John Wiley & Sons, 2012.

JACOBS, Adam.: The pathologies of big data. Communications of the ACM, 2009, 52.8: 36-44.

JANG, Haeng-Jin, et al.: Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media. Expert Systems with applications, 2013, 40.18: 7492-7503.

JANSEN, Bernard J., et al.: Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology, 2009, 60.11: 2169-2188.

JIFA, Gu; LINGLING, Zhang.: Data, DIKW, Big data and Data science. Procedia Computer Science, 2014, 31: 814-821.

JIN, Xiaolong, et al.: Significance and challenges of big data research. Big Data Research, 2015, 2.2: 59-64.

JOHNSON, Jeanne E.: Big data+ big analytics= big opportunity: big data is dominating the strategy discussion for many financial executives. As these market dynamics continue to evolve, expectations will continue to shift about what should be disclosed, when and to whom. Financial Executive, 2012, 28.6: 50-54.

JOSHI, Mahesh, et al.: Movie reviews and revenues: An experiment in text regression. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. p. 293-296.

KALOGIROU, Soteris A.: Applications of artificial neural-networks for energy systems. Applied energy, 2000, 67.1: 17-35.

KANG, Daekook; PARK, Yongtae.: Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. Expert Systems with Applications, 2014, 41.4: 1041-1050.

KAPLAN, Andreas M.; HAENLEIN, Michael.: Users of the world, unite! The challenges and opportunities of Social Media. Business horizons, 2010, 53.1: 59-68.

KARYPIS, George; HAN, Eui-Hong; KUMAR, Vipin.: Chameleon: Hierarchical clustering using dynamic modeling. Computer, 1999, 32.8: 68-75.

KAYAHAN, Dilek; SERGIN, Aslı; DIRI, Banu.: Determination of TV programme ratings by twitter. In: Signal Processing and Communications Applications Conference (SIU), 2013 21st. IEEE, 2013. p. 1-4.

KHUN, Max; WESTON, Steve; COULTER, Nathan; CULP, Mark.: Package ‘C50’. 2015.

KIETZMANN, Jan H., et al.: Social media? Get serious! Understanding the functional building blocks of social media. Business horizons, 2011, 54.3: 241-251.

KIM, Taegu; HONG, Jungsik; KANG, Pilsung.: Box office forecasting using machine learning algorithms based on SNS data. International Journal of Forecasting, 2015, 31.2: 364-390.

LIU, Bing.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, Second Edition. Chapman and Hall/CRC, 2010. p. 627-666.

LIU, Bing.: Sentiment analysis and opinion mining: Synthesis lectures on human language technologies [M].[sl]: Morgan & Claypool Publishers, 2012: 1–167. Google Scholar.

LIU, Hugo.: Social network profiles as taste performances. Journal of Computer-Mediated Communication, 2007, 13.1: 252-275.

MACQUEEN, James, et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967. p. 281-297.

MANYIKA, James, et al.: Big data: The next frontier for innovation, competition, and productivity. 2011.

MCCLELLAND, James L.; RUMELHART, David E.; HINTON, Geoffrey E.: The appeal of parallel distributed processing. MIT Press, Cambridge MA, 1986, 3-44.

MCCULLOCH, Warren S.; PITTS, Walter.: A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 1943, 5.4: 115-133.

MEDHAT, Walaa; HASSAN, Ahmed; KORASHY, Hoda.: Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 2014, 5.4: 1093-1113.

MERAL, Meric; DIRI, Banu.: Sentiment analysis on Twitter. In: Signal Processing and Communications Applications Conference (SIU), 2014 22nd. IEEE, 2014. 690-693.

MEYER, David; DIMITRIADOU, Evgenia; HORNIK, Kurt; WEINGESSEL, Andreas; LEISCH, Friedrich; CHANG, Chih-Chung; LIN, Chih-Chen.: e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-8. 2017.

MICHEL, Frank.: "How Many Photos Are Uploaded to Flickr Every Day and Month?" http://www.flickr.com/photos /franckmichel/6855169886/, 2012.

MISHNE, Gilad, et al.: Predicting Movie Sales from Blogger Sentiment. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006. p. 155-158.

MISLOVE, Alan, et al.: Understanding the Demographics of Twitter Users. ICWSM, 2011, 11: 5th.

MITCHELL, Lewis; DODDS, Peter-Sheridan et al.: The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. PloS one, 2013, 8.5: e64417.

MITTAL, Anshul; GOEL, Arpit.: Stock prediction using twitter sentiment analysis. Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), 2012, 15.

MOORE, Luke.: Fourth Source. Is your Advertising Campaign Driving Intent to Purchase? , 2014

MOSTAFA, Mohamed M.: More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications, 2013, 40.10: 4241-4251.

MUKKAMALA, Raghava Rao; VATRAPU, Towards a formal model of social data. IT-Universitetet i København, 2013.

Ravi; HUSSAIN, Abid.:

NAKIP, Mahir.: Pazarlama araştırmaları teknikler ve (SPSS destekli) uygulamalar. Seçkin Yayıncılık, 2003.

NEMSCHOFF, Michele.: Social Media Marketing: How Big Data is Changing Everything. CMS Wire, 2013, 16.

NEVIAROUSKAYA, Alena; PRENDINGER, Helmut; ISHIZUKA, Mitsuru.: Textual affect sensing for sociable and expressive online communication. Affective Computing and Intelligent Interaction, 2007, 218-229.

OH, Chong; YERGEAU, Stephanie.: Social capital, social media, and TV ratings. International Journal of Business Information Systems, 2017, 24.2: 242-260.

OHSUMI, Noboru.: From data analysis to data science. In: Data Analysis, Classification, and Related Methods. Springer Berlin Heidelberg, 2000. p. 329-334.

OLSON, David L.; WU, Desheng.: Predictive Data Mining Models. Computational Risk Management, 2016.

ÖZDAĞOĞLU, Güzin; KAPUCUGIL-İKIZ, Aysun; ÇELIK, Ayhan Fuat.: Topic modelling-based decision framework for analysing digital voice of the customer. Total Quality Management & Business Excellence, 2016, 1-18.

PANG, Bo; LEE, Lillian.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005. p. 115-124.

PENTLAND, Alex.: Reinventing society in the wake of big data. Edge. org Conversation, August, 2012.

PITTMAN, Matthew; TEFERTILLER, Alec C.: With or without you: Connected viewing and co-viewing Twitter activity for traditional appointment and asynchronous broadcast television models. First Monday, 2015, 20.7.

PORSHNEV, Alexander; REDKIN, Ilya; SHEVCHENKO, Alexey.: Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. In: Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. IEEE, 2013. p. 440-444.

PROVOST, Foster; FAWCETT, Tom.: Data science and its relationship to big data and data-driven decision making. Big Data, 2013, 1.1: 51-59.

RAHMAN, Mohammad Naimur; ESMAILPOUR, Amir; ZHAO, Junhui.: Machine Learning with Big Data An Efficient Electricity Generation Forecasting System. Big Data Research, 2016, 5: 9-15.

ROSENBLATT, Frank.: A model for experiential storage in neural networks. Computer and information sciences. Washington, DC: Spartan, 1964.

RUI, Huaxia; LIU, Yizao; WHINSTON, Andrew.: Whose and what chatter matters? The effect of tweets on movie sales. Decision Support Systems, 2013, 55.4: 863-870.

RUSSOM, Philip, et al.: Big data analytics. TDWI best practices report, fourth quarter, 2011, 19: 40.

SAGIROGLU, Seref; SINANC, Duygu.: Big data: A review. In: Collaboration Technologies and Systems (CTS), 2013 International Conference on. IEEE, 2013. p. 42-47.

SAKAKI, Takeshi; OKAZAKI, Makoto; MATSUO, Yutaka.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web. ACM, 2010. p. 851-860.

SAN, Ohn Mar; HUYNH, Van-Nam; NAKAMORI, Yoshiteru.: An alternative extension of the k-means algorithm for clustering categorical data. International Journal of Applied Mathematics and Computer Science, 2004, 14.2: 241-248.

SCHLAGWEIN, Daniel.: Strategic tools: how firms successfully use social media. from http://www.smartcompany.com.au/leadership/management/41115-strategic-tools-how-firms-successfully-use-social-media.html, 2014

SCHNEIDER, R. D.: Custom Hadoop for Dummies, Special Edition. John Wiley & Sons Incorporated, 2012.

SHARDA, Ramesh; DELEN, Dursun.: Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications, 2006, 30.2: 243-254.

SILVA, Thiago H., et al.: A picture of Instagram is worth more than a thousand words: Workload characterization and application. In: Distributed Computing in Sensor Systems (DCOSS), IEEE International Conference on. IEEE, 2013. p. 123-132.

SOMAN, K. P.; LOGANATHAN, R.; AJAY, V.: Machine learning with SVM and other kernel methods. PHI Learning Pvt. Ltd., 2009.

STEPHEN, Andrew T.; TOUBIA, Olivier.: Deriving value from social commerce networks. Journal of marketing research, 2010, 47.2: 215-228.

STIEGLITZ, Stefan, et al.: Social media analytics. Wirtschaftsinformatik, 2014, 56.2: 101-109.

ŞİMŞEK, Tuğba U.; Veri madenciliği ve müşteri ilişkileri yönetiminde (CRM) bir uygulama, 2006.

ŞİMŞEK, Mehmet; ÖZDEMİR, Ulvi.; Analysis of the relation between Turkish twitter messages and stock market index. In: Application of Information and Communication Technologies (AICT), 6th International Conference on. IEEE, 2012. p. 1-4.

TAN, Pang-Ning; STEINBACH, Micahel; KUMAR, Vipin.: Introduction to Data Mining. Person Education. Inc., New Delhi, 2006.

THELWALL, Mike, et al.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 2010, 61.12: 2544-2558.

TRATTNER, Christoph; KAPPE, Frank.: Social stream marketing on Facebook: a case study. International Journal of Social and Humanistic Computing, 2013, 2.1-2: 86-103.

ÜNAL, Fatih.: Büyük veri ve semantik. Abaküs, 2015.

VAPNIK, Vladimir; CHERVONENKIS, Alexey.: A note on one class of perceptrons. Automation and remote control, 1964, 25.1: 103.

VASANT, D.: Data Science and Prediction. Communications of the ACM. 2013.

WAKAMIYA, Shoko; LEE, Ryong; SUMIYA, Kazutoshi.: Crowd-powered TV viewing rates: measuring relevancy between tweets and TV programs. In: International Conference on Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2011. p. 390-401.

WAKAMIYA, Shoko; RYONG, L. E. E.; SUMIYA, Kazutoshi.: Twitter-based TV Audience Behavior Estimation for Better TV Ratings. In: DEIM Forum (http://db-event. jpn. org/deim2011/). 2011.

WAMBA, Samuel Fosso, et al.: The Primer of Social Media Analytics. Journal of Organizational and End User Computing (JOEUC), 2016, 28.2: 1-12.

WEICHSELBRAUN, Albert; GINDL, Stefan; SCHARL, Arno.: Enriching semantic knowledge bases for opinion mining in big data applications. Knowledge-Based Systems, 2014, 69: 78-85.

WU, Xindong, et al.:	Data mining with big data. ieee transactions on knowledge and data engineering, 2014, 26.1: 97-107.
XIA, Rui; ZONG, Chengqing; LI, Shoushan.:	Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 2011, 181.6: 1138-1152.
XIANG, Zheng, et al.:	What can big data and text analytics tell us about hotel guest experience and satisfaction?. International Journal of Hospitality Management, 2015, 44: 120-130.
YOUNG, Sean D.:	A "big data" approach to HIV epidemiology and prevention. Preventive medicine, 2015, 70: 17-18.
ZAILSKAITE-JAKSTE, Ligita; KUVYKAITE, Rita.:	Consumer engagement in social media by building the brand. In: Proceedings in EIIC-1st Electronic International Interdisciplinary Conference. 2012.
ZHANG, Li; LUO, Jianhua; YANG, Suying.:	Forecasting box office revenue of movies with BP neural network. Expert Systems with Applications, 2009, 36.3: 6580-6587.
ZHANG, Linhao.:	Sentiment analysis on Twitter with stock price and significant keyword correlation. 2013. PhD Thesis.
ZHANG, Xue; FUEHRES, Hauke; GLOOR, Peter A.:	Predicting stock market indicators through twitter "I hope it is not as bad as I fear". Procedia-Social and Behavioral Sciences, 2011, 26: 55-62.
ZHAO, Wayne Xin, et al.:	Comparing twitter and traditional media using topic models. In: European Conference on Information Retrieval. Springer Berlin Heidelberg, 2011. p. 338-349.
ZHENG, Xiaolin; ZHU, Shuai; LIN, Zhangxi.:	Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach. Decision Support Systems, 2013, 56: 211-222.
ZIKOPOULOS, Paul, et al.:	Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

Elektronik Kaynaklar

Github:	https://github.com/Jefferson-Henrique/GetOldTweets-python
Habertürk:	http://hbrturkiye.com/video/vaka-ii-yemek-sepeti-ve-buyuk-veri
Netflix Prize:	http://www.netflixprize.com//community/viewtopic.php? id=1537; accessed July 9, 2012
Python:	Python, 2017. https://www.python.org/about/
Rapidminer:	Rapidminer, 2017. https://rapidminer.com
R-Project:	https://cran.r-project.org/web/packages/RSNNS/RSNNS.pdf
R-Project:	https://cran.r-project.org/web/packages/e1071/e1071.pdf
R-Project:	R-Project, 2017. https://www.r-project.org/
R-Project:	https://cran.r-project.org/web/packages/C50/C50.pdf
Twitter:	Twitter, API Rate Limits, 2015. https://dev.twitter.com/rest/public/rate-limiting

ÖZGEÇMİŞ

BURCU KARAÖZ

Öğrenim Bilgisi

Doktora 2014- Devam ediyor	İSTANBUL ÜNİVERSİTESİ SOSYAL BİLİMLER ENSTİTÜSÜ/SAYISAL YÖNTEMLER (DR) Tez adı: Büyük Veri Üzerinde Veri Madenciliği Yöntemleri: Bir Araştırma Tez Danışmanı:(UMMAN TUĞBA ŞİMŞEK GÜRSOY)
Yüksek Lisans 2012-2014	HACETTEPE ÜNİVERSİTESİ İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ/İŞLETME BÖLÜMÜ/SAYISAL YÖNTEMLER ANABİLİM DALI Tez adı: Maden Üretim Planlaması Ve Çizelgelemesi Üzerine Bir Tam Sayılı Programlama Önerisi: Kar Maden Örneği Tez Danışmanı:(ŞAHAP ARMAĞAN TARIM)
Lisans 2007-2012	HACETTEPE ÜNİVERSİTESİ İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ/İŞLETME BÖLÜMÜ/İŞLETME PR. (İNGİLİZCE)

Görevler

ARAŞTIRMA GÖREVLİSİ 2013- Devam ediyor	YAŞAR ÜNİVERSİTESİ/İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ/İŞLETME BÖLÜMÜ/İŞLETME PR. (İNGİLİZCE) (ÜCRETLİ))

Projelerde Yaptığı Görevler:

1 İZMİR TİCARET BORSASI İŞ ZEKASI VE BÜYÜK VERİ ANALİZİ PROJESİ, Yükseköğretim Kurumları tarafından destekli bilimsel araştırma projesi, Araştırmacı, 06/04/2016 – 06/10/2016 (ULUSAL)

2 SÜREÇ İYİLEŞTİRME VE YÖNETİM ORGANİZASYON PROJESİ, Özel Kuruluşlar, Araştırmacı, 23/06/2014 - 23/06/2015 (ULUSAL)

Eserler

A. Uluslararası hakemli dergilerde yayımlanan makaleler:

1 KOCAMAN YELİZ,KARAÖZ BURCU,DİZBAY İKBAL ECE,GÜMÜŞOĞLU ŞEVKİNAZ (2017). DEĞİŞİM YÖNETİMİNDE İŞ GÜCÜ PLANLAMA YAKLAŞIMI BİR BÜYÜKŞEHİR BELEDİYESİ UYGULAMASI Uluslararası İktisadi ve İdari İncelemeler Dergisi, (), 515-526.

2 KARAÖZ BURCU (2016). REDESIGNING THE ASSIGNMENT OF INTERNAL AUDITORS BASED ON AUDITOR UTILITY. Business And Management Studies: An International Journal, 4(3), 246-259., Doi: 10.15295/bmij.v4i3.159 (Yayın No: 2995381)

B. Uluslararası bilimsel toplantılarda sunulan ve bildiri kitaplarında (proceedings) basılan bildiriler :

1 GÜMÜŞOĞLU ŞEVKİNAZ, KARAÖZ BURCU (2014) Planning Capacity Requirements and Finishing Time of a

Packing Factory Building or Buying Decision Project. 4th Conference of the International Network of Business and Management Journals (INBAM) (Tam Metin Bildiri/)(Yayın No:1325379)

2 GÜMÜŞOĞLU ŞEVKİNAZ, KARAÖZ BURCU (2014) Project Crashing and Risk Based Resource Allocation A Holiday Village Example. 9th Annual London Business Research Conference (Tam Metin Bildiri/)(Yayın No:1325375)

3 AKIN GÖKTÜRK, KARAÖZ BURCU (2015). An outlook on Turkish Private Equity and Venture Capital Market. 2nd International Borsa Istanbul Finance and Economics Conference (Özet Bildiri/)(Yayın No:2813694)

4 GÜMÜŞOĞLU ŞEVKİNAZ, KARAÖZ BURCU (2013). Tarihsel Surecte Girisimcilik Muğla Örneği. ULUSLARARASI GİRİŞİMCİLİK VE KARİYER SEMPOZYUMU (Tam Metin Bildiri/)(Yayın No:607552)

D. Ulusal hakemli dergilerde yayımlanan makaleler :

1 GÜMÜŞOĞLU ŞEVKİNAZ, KARAÖZ BURCU (2014). Tarihsel Süreçte Girişimcilik Muğla Örneği. Ekonomi ve Yönetim Araştırmaları Dergisi, 3(1) (Kontrol No: 1325388)

E. Ulusal bilimsel toplantılarda sunulan ve bildiri kitaplarında basılan bildiriler:

1 KARAÖZ BURCU,KILIÇ ONUR ALPER,TARIM ŞAHAP ARMAĞAN (2015). Maden Üretimi Planlama ve Çizelgeleme Üzerine Bir Tam Sayili Programlama Uygulaması. YAEM, ODTU (Özet Bildiri/)(Yayın No:2813685)

2 DİZBAY İKBAL ECE,KOCAMAN YELİZ,KARAÖZ BURCU (2015). Vikor Yöntemi ile Kamu Kuruluşlarından Yardım Alacak Kişilerin Seçimi ve Dağıtım Rotalarının Belirlenmesi. YAEM, ODTU (Özet Bildiri/)(Yayın No:2813690)

EKLER

Ek 1: Reyting ölçümünde kullanılan Sosyoekonomik statü grupları

Kriterler;

- Meslek
- Eğitim
- Gelir Seviyesi
- Gelirin kaynağı
- Yaşanılan evin tipi
- Yaşanılan çevrenin yapısı
- Sahip olunan mal, mülk (ev, yazlık, otomobil, elektronik eşya)

A SES Grubu: Yüzde 4

- Hemen hepsi üniversite mezunu, yüzde 30 dolayında lisansüstü.
- Yarıya yakını, ücretli çalışan, nitelikli uzman (avukat, doktor, mühendis vb.)...
- Yüzde 10'a yakını, 20'den fazla çalışanı olan beyaz yakalı.
- Yüzde 20'si irili ufaklı işyeri sahibi (bunların yarıya yakınının yanında çalışanı yok)...
- Eşi olan AGG'lerin yüzde 40'a yakınının eşi çalışıyor.
- Hanelerin yüzde 20'si para biriktiriyor.
- Yüzde 30'u tatilini tatil köyü/otele giderek değerlendiriyor.
- Hanelerin yarısına yakınında kitaplık/kütüphane var.

B SES Grubu: Yüzde 9

- Üniversite/lisansüstü oranı yüzde 60'larda... Yüzde 35 civarında 2 yıllık yüksek okul veya lise mezunu.
- Yüzde 60'ı memur, teknik personel, uzman (yönetici olmayan)...
- Yüzde 15'i irili ufaklı işyeri sahibi (bunların çoğunun yanında 1-5 arası çalışanı var)...
- Eşi olan AGG'lerin yüzde 30'unun eşi çalışıyor.
- Hanelerin yüzde 13'ü para biriktiriyor.
- Yüzde 20'si tatilini tatil köyü/otele giderek değerlendiriyor.
- Hanelerin yüzde 30'unda kitaplık/kütüphane var.

C1 SES Grubu: Yüzde 22

- Yüzde 60'ı lise mezunu (Bunun içinde yüzde 20 meslek lisesi); yüzde 10'u yüksekokul ve üstü.
- Yüzde 40'ı esnaf, dükkan sahibi; yüzde 30'u kalifiye işçi (lise eğitimli)...
- Yüzde 15'e yakın memur, teknik eleman.
- Yüzde 15'e yakını emekli.

- Eşi olan AGG'lerin yüzde 13'ünün eşi çalışıyor.
- Hanelerin yüzde 5'i para biriktiriyor.
- Yüzde 20'si tatilini tatil köyü/otele giderek, yüzde 40'a yakını yakınlarını ziyaret ederek değerlendiriyor.
- Hanelerin yüzde 20'ye yakınında kitaplık/kütüphane var.

C2 SES Grubu: Yüzde 29

- Yüzde 20'ye yakını lise mezunu... Ortaokul ve daha düşük eğitimli oranı yüzde 80.
- Çoğunlukla ilkokul mezunu, düzenli çalışan işçi (Yüzde 60'lar dolayında)...
- Yüzde 10 kadarı tek başına seyyar olarak çalışıyor.
- Yüzde 20'si emekli, çalışmıyor.
- Eşi olan AGG'lerin eşinin çalışma oranı yüzde 10'un altında.
- Yüzde 70'i tatile çıkmıyor, çıkanlar yakınlarını ziyaret etmek için memlekete gidiyorlar (yüzde 25)...
- Hanelerin yüzde 10'unda kitaplık/kütüphane var.

D SES Grubu: Yüzde 28

- Yüzde 70'in üzerinde ilkokul mezunu veya illkokul terk, gerisi ortaokul.
- Yüzde 30 kadar emekli, çalışmıyor.
- Yüzde 20'nin üzerinde işçi (çoğunlukla parça-başı çaışan)...
- Yüzde 30'u küçük çaplı çiftçi.
- Yüzde 10'a yakını ev kadını.
- Yüzde 80'i tatile çıkmıyor, gerisi memlekete gidiyor.

E SES Grubu: Yüzde 9

- Yüzde 95'i ilkokul mezunu veya ilkokul terk.
- Yüzde 30'a yakını işsiz (çoğu yardımla geçiniyor)
- Yüzde 40'ı emekli, çalışmıyor; yüzde 30'u emekli, işçi olarak çalışıyor.
- Geri kalan yüzde 20'nin üzerinde hanede AGG ev-kadını (düzenli geliri olmayan yardımla geçinen)...

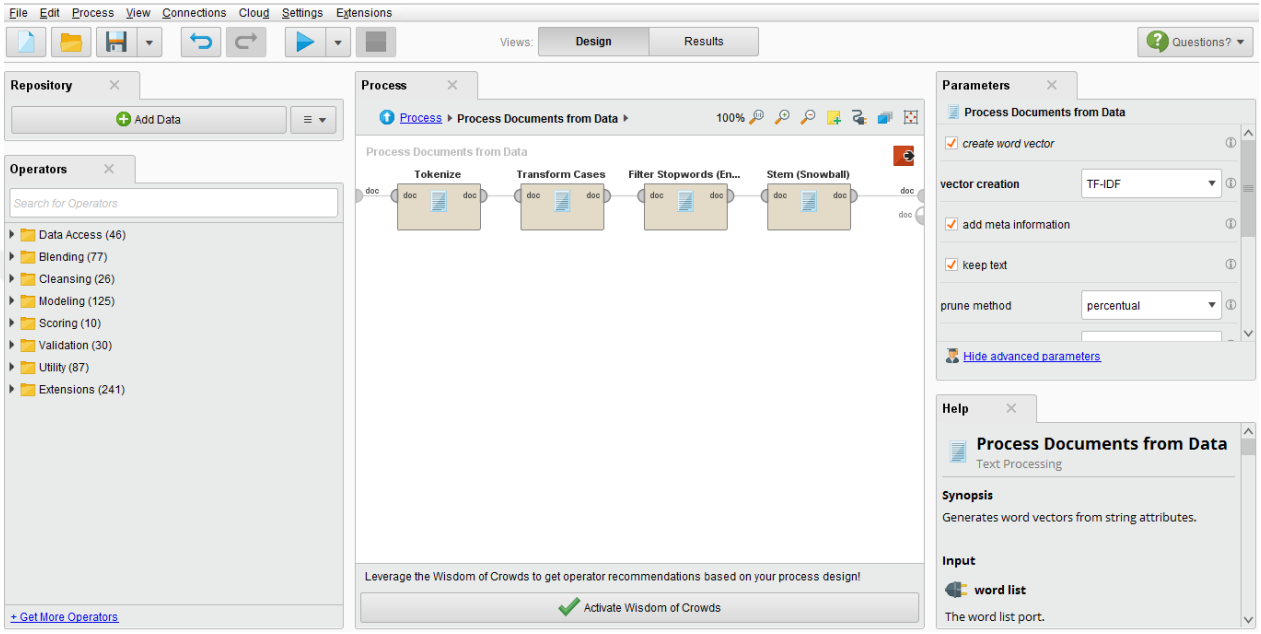
Kaynak: http://kitleiletisimi.blogspot.com.tr/2013/06/turkiyenin-sosyo-ekonomik-statu-ses.html

Ek 2: Durak kelimeler

acaba	birşey	diye	hangi	kendisi	olduğu	oysa	tam
altı	biz	doksan	hangisi	kendisine	olduğunu	pek	tamam
altmış	bizden	dokuz	hani	kendisini	olduklarını	rağmen	tamamen
ama	bize	dolayı	hatta	kez	olmadı	sana	tarafından
ancak	bizi	dolayısıyla	hem	ki	olmadığı	sanki	trilyon
arada	bizim	dört	henüz	kim	olmak	sanki	tüm
artık	böyle	e	hep	kime	olması	şayet	tümü
asla	böylece	edecek	hepsi	kimi	olmayan	şekilde	üç
aslında	bu	eden	her	kimin	olmaz	sekiz	üzere
aslında	buna	ederek	herhangi	kimisi	olsa	seksen	var
ayrıca	bunda	edilecek	herkes	kimse	olsun	sen	vardı
az	bundan	ediliyor	herkese	kırk	olup	senden	ve
bana	bunlar	edilmesi	herkesi	madem	olur	seni	veya
bazen	bunları	ediyor	herkesin	mi	olur	senin	ya
bazı	bunların	eğer	hiç	mı	olursa	şey	yani
bazıları	bunu	elbette	hiçbir	milyar	oluyor	şeyden	yapacak
belki	bunun	elli	hiçbiri	milyon	on	şeye	yapılan
ben	burada	en	için	mu	ön	şeyi	yapılması
benden	bütün	etmesi	içinde	mü	ona	şeyler	yapıyor
beni	çoğu	etti	iki	nasıl	önce	şimdi	yapmak
benim	çoğunu	ettiği	ile	ne	ondan	siz	yaptı
beri	çok	ettiğini	ilgili	neden	onlar	sizden	yaptığı
beş	çünkü	fakat	ise	nedenle	onlara	size	yaptığını
bile	da	falan	işte	nerde	onlardan	sizi	yaptıkları
bilhassa	daha	filan	itibaren	nerede	onları	sizin	yedi
bin	dahi	gene	itibariyle	nereye	onların	sonra	yerine
bir	dan	gereği	kaç	neyse	onu	şöyle	yetmiş
biraz	de	gerek	kadar	niçin	onun	şu	yine
birçoğu	defa	gibi	karşın	niye	orada	şuna	yirmi
birçok	değil	göre	kendi	öbür	öte	şunları	yoksa
biri	diğer	hala	kendilerine	olan	ötürü	şunu	yüz
birisi	diğeri	halde	kendine	olarak	otuz	ta	zaten
birkaç	diğerleri	halen	kendini	oldu	öyle	tabii	zira

Ek 3: Metin ön işleme aşaması- uygulama

Rapidminer programı ile metin ön işleme



R ile metin ön işleme

```
install.packages(“tm”)
library(tm)
setwd("C:\\location")
text<-readLines("test.txt")
corpus<-Corpus(VectorSource(text))
length(corpus)
corpus[[1]]
tdm = TermDocumentMatrix(corpus, control = list(weighting = weightTfIdf, removePunctuation = T, removeNumbers = T, stopwords= T, stemming = T))
dtmMatrix <- as.matrix(dtm)
write.csv(tdmMatrix, 'mytestfile.csv')
freq<-slam::row_sums(tdm, na.rm = T)
high.freq=tail(sort(freq),n=1000)
```

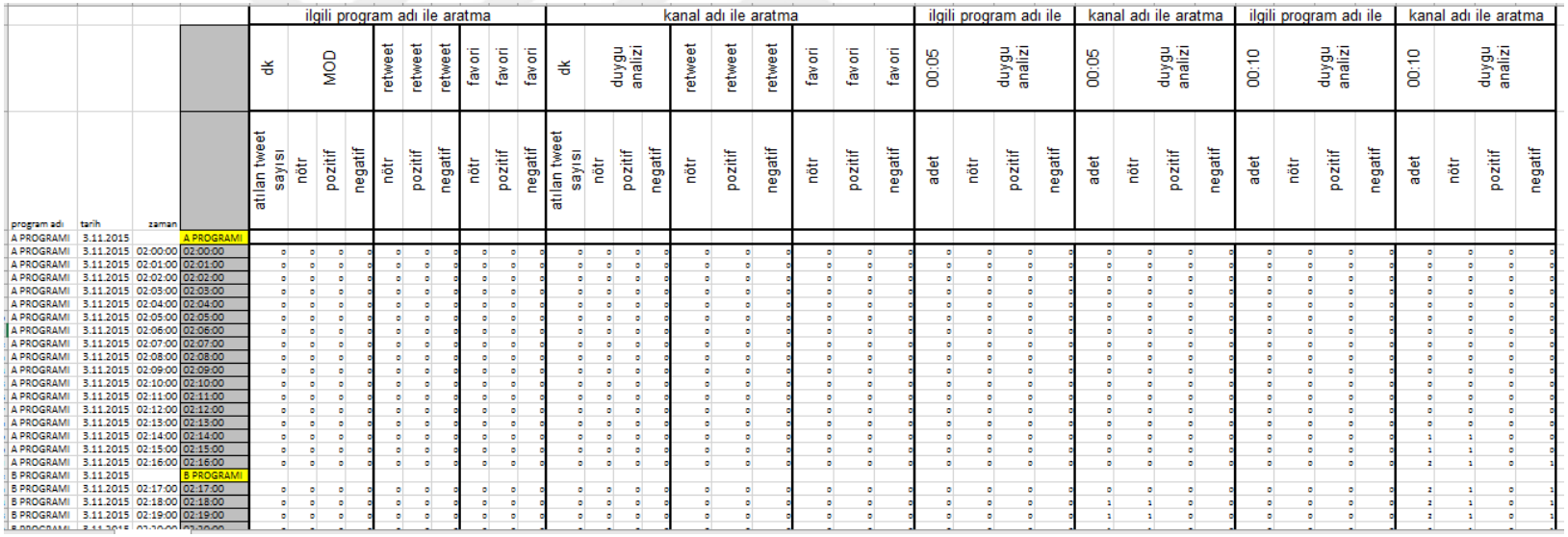
Terim doküman matris örneği

ağrı	çok	gülmek	karın	komik	inanılmaz	sempatik	text
0.548	0	0.625	0.512	0	0	0	karnım ağrıdı gülmekten
0	0.695	0	0	0.614	0	0	çok komik ya
0	0	0	0	0	0.520	0.548	Ya inanılmaz sempatik

Ek 4: Excel fonksiyonları ve VBA kodları

```
Sub duyguanalizi()
'
' duyguanalizi Macro
'
' Keyboard Shortcut: Ctrl+b
'
    Range("E3").Select
    ActiveCell.FormulaR1C1 = "=IF(R2C="""",0,IF(ISNUMBER(SEARCH(R2C,RC4)),
1,0))"
    Range("E3").Select
    Selection.AutoFill Destination:=Range("E3:MQ3"), Type:=xlFillDefault
    Range("E3:MQ3").Select
    Selection.AutoFill Destination:=Range("E3:MQ20001")
    Range("E3:MQ20001").Select
    Range("D3").Select
    Selection.End(xlToLeft).Select
    ActiveCell.FormulaR1C1 = "=SUM(RC[4]:RC[199])"
    Range("B3").Select
    ActiveCell.FormulaR1C1 = "=SUM(RC[199]:RC[353])"
    Range("C3").Select
    ActiveCell.FormulaR1C1 = _
        "=IF(AND(RC[-2]=0,RC[-1]=0),0,IF(AND(RC[-2]=0,RC[-1]<>0),2,IF(AND(RC[-
1]=0,RC[-2]<>0),1,IF(RC[-2]>RC[-1],1,IF(RC[-2]=RC[-1],0,2)))))"
    Range("A3:C3").Select
    Selection.AutoFill Destination:=Range("A3:C20001")
    Range("A3:C20001").Select
End Sub
```

Ek 5: Dakikalık reyting veri seti örnek görseli



				ilgili program adı ile aratma									kanal adı ile aratma									ilgili program adı ile				kanal adı ile aratma				ilgili program adı ile				kanal adı ile aratma			
				dk	MOD			retweet	retweet	retweet	favori	favori	favori	dk	duygu analizi			retweet	retweet	retweet	favori	favori	favori	00:05	duygu analizi			00:05	duygu analizi			00:10	duygu analizi			00:10	duygu analizi		
program adi	tarih	zaman		atılan tweet sayısı	nötr	pozitif	negatif	nötr	pozitif	negatif	nötr	pozitif	negatif	atılan tweet sayısı	nötr	pozitif	negatif	nötr	pozitif	negatif	nötr	pozitif	negatif	adet	nötr	pozitif	negatif	adet	nötr	pozitif	negatif	adet	nötr	pozitif	negatif	adet	nötr	pozitif	negatif
A PROGRAMI	3.11.2015		A PROGRAMI																																		
A PROGRAMI	3.11.2015	02:00:00	02:00:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:01:00	02:01:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:02:00	02:02:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:03:00	02:03:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:04:00	02:04:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:05:00	02:05:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:06:00	02:06:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:07:00	02:07:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:08:00	02:08:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:09:00	02:09:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:10:00	02:10:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:11:00	02:11:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:12:00	02:12:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:13:00	02:13:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:14:00	02:14:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:15:00	02:15:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A PROGRAMI	3.11.2015	02:16:00	02:16:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B PROGRAMI	3.11.2015		B PROGRAMI																																		
B PROGRAMI	3.11.2015	02:17:00	02:17:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1
B PROGRAMI	3.11.2015	02:18:00	02:18:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	1
B PROGRAMI	3.11.2015	02:19:00	02:19:00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	1

Ek 6: Program reyting veri seti örnek görseli

program adı	tarih	program başlangıç	program bitiş	Share: Individuals 5+	Share: Individuals SES AB	Share: Individuals 20+ ABC1	Rtg%: Individuals 5+	Rtg%: Individuals SES AB	Rtg%: Individuals 20+ ABC1	program adı ile atılan toplam tweet sayısı	nötr	pozitif	negatif	program adı ile: retweet nötr	program adı ile: retweet pozitif	program adı ile: retweet negatif	program adı ile: favori nötr	program adı ile: favori pozitif	program adı ile: favori negatif	kanal adı ile atılan toplam tweet sayısı	nötr	pozitif	negatif	kanal adı ile: retweet nötr	kanal adı ile: retweet pozitif	kanal adı ile: retweet negatif	kanal adı ile: favori nötr	kanal adı ile: favori pozitif	kanal adı ile: favori negatif	program adı ile: Program süresi +-2 de atılan toplam tweet
A	3.11.2015	02:00:00	02:16:00	6.69	1.22	6.69	0.43	0.07	0.47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
B	3.11.2015	02:17:00	03:06:00	5.03	2.34	4.19	0.24	0.09	0.21	0	0	0	0	0	0	0	0	0	0	5	3	1	1	0	0	1	4	1	0	0
C	3.11.2015	03:23:00	04:20:00	2.92	1.40	1.38	0.09	0.03	0.04	0	0	0	0	0	0	0	0	0	0	3	1	2	0	0	1	0	5	0	0	0
D	3.11.2015	04:21:00	06:15:00	2.78	1.88	4.00	0.06	0.03	0.08	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0
E	3.11.2015	06:17:00	06:58:00	4.48	1.97	3.36	0.13	0.05	0.08	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	8	0	0
F	3.11.2015	07:00:00	07:03:00	2.64	0.67	1.42	0.11	0.02	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	3.11.2015	07:04:00	07:24:00	1.39	0.94	1.06	0.07	0.04	0.04	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
B	3.11.2015	07:38:00	08:40:00	1.36	3.54	2.52	0.10	0.18	0.16	0	0	0	0	0	0	0	0	0	0	4	1	3	0	18	8	0	48	5	0	0
H	3.11.2015	08:41:00	08:59:00	1.61	2.36	2.73	0.15	0.16	0.23	0	0	0	0	0	0	0	0	0	0	3	2	1	0	0	0	0	0	0	0	0
D	3.11.2015	09:00:00	12:18:00	2.25	5.28	3.32	0.30	0.53	0.42	10	4	3	3	6	1	1	32	1	2	43	25	13	5	64	9	4	130	33	2	10
D	3.11.2015	12:19:00	13:18:00	1.83	3.49	2.40	0.26	0.40	0.33	0	0	0	0	0	0	0	0	0	0	9	6	3	0	6	5	0	74	49	0	0
C	3.11.2015	13:19:00	13:38:00	2.29	3.03	1.99	0.30	0.27	0.24	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	34	0	0	69	0
C	3.11.2015	13:39:00	15:01:00	2.30	5.05	3.18	0.29	0.42	0.36	0	0	0	0	0	0	0	0	0	0	18	11	6	1	33	10	0	164	48	0	1
A	3.11.2015	15:02:00	15:15:00	1.54	4.78	3.05	0.20	0.39	0.34	1	1	0	0	0	0	0	0	0	0	3	3	0	0	1	0	0	1	0	0	12
A	3.11.2015	15:16:00	18:49:00	5.67	8.53	8.08	1.11	1.21	1.29	8	6	0	2	0	0	0	1	0	3	64	40	16	8	121	21	40	295	78	68	20
B	3.11.2015	18:50:00	18:56:00	3.01	2.21	3.13	0.97	0.63	0.86	0	0	0	0	0	0	0	0	0	0	2	1	1	0	2	0	0	30	0	0	0

Ek 7: Karar ağacı algoritması uygulama kod ve sonuçlar

```
install.packages("C50") install.packages("openxlsx")
library(openxlsx) library(C50)
setwd("C:\\Users\\.....")
dat <- read.xlsx("veriseti.xlsx", sheet=1)
#sınıf değişkenini düzenle
dat$y     <-factor(dat$y)
#veriyi test ve train diye ayir
index <- 1:nrow(dat)
testindex <- sample(index, trunc(length(index)/5))
testset <- dat[testindex,]
trainset <- dat[-testindex,]
#karar ağacı modelini kur
treeModel <- C5.0(x = trainset[, -28], y = trainset$y)
#önemli değişkenleri bul
C5imp(treeModel,metric = "splits")
summary(treeModel)
#kural modeli kur
ruleModel <- C5.0(y ~ ., data = trainset, rules = TRUE)
summary(ruleModel)
#tahmin
c50pred <- predict(treeModel, testset[-28])
summary(c50pred)
#karşılaştırma matrisi
table(predict=c50pred, truth=testset$y)
#doğruluk oranı
acc=mean(c50pred==testset$y)
acc
```

5+ very seti için sonuçlar

```
>treeModel <- C5.0(x = trainset[, -28], y = trainset$y)
>C5imp(treeModel,metric = "splits")
```

	Overall		
x13	10.0840336	x18	2.5210084
x3	10.0840336	x20	2.5210084
x1	9.2436975	x26	2.5210084
x11	9.2436975	x29	2.5210084
x7	7.5630252	x30	2.5210084
x31	5.8823529	x19	1.6806723
x17	4.2016807	x21	1.6806723
x28	4.2016807	x22	1.6806723
x14	3.3613445	x23	1.6806723
x2	3.3613445	x27	1.6806723
x24	3.3613445	x12	0.8403361
x4	3.3613445	x16	0.8403361
x15	2.5210084	x25	0.8403361
		x6	0.0000000

```
> summary(treeModel)
Call:C5.0.default(x = trainset[, -28], y = trainset$y)

C5.0 [Release 2.07 GPL Edition]        Fri Apr 07 17:53:00 2017

-------------------------------

Class specified by attribute `outcome'

Read 2460 cases (28 attributes) from undefined.data

Decision tree:

x2 = 1:

:...x1 in {2,3}:

:   :...x4 = 0: c (3/1)

:   :   x4 = 1:

:   :   :...x1 = 3: b (12)
```

: : x1 = 2:
: : :...x3 in {21,22,31,32,36,43,45,51,61,71,72,77,78,79,
: : : 87}: c (9)
: : x3 = 76: b (10/4)
: x1 = 1:
: :...x3 in {21,22,31,32,36,43,45,51,61,71,72}: e (0)
: x3 = 78: d (19/1)
: x3 in {79,87}: c (6/1)
: x3 = 77:
: :...x24 > 46:
: : :...x19 <= 201: e (3)
: : : x19 > 201: d (2)
: : x24 <= 46:
: : :...x21 > 0: c (2)
: : x21 <= 0:
: : :...x14 <= 3: c (2)
: : x14 > 3: d (6/1)
: x3 = 76:
: :...x28 <= 1:
: :...x7 in {3,4}: c (8/3)
: : x7 = 6: e (14/1)
: : x7 = 1:
: : :...x11 <= 1793: e (6/1)
: : : x11 > 1793:
: : : :...x11 <= 4457: c (2)
: : : x11 > 4457: e (7/1)
: : x7 = 5:
: : :...x11 <= 663: d (2) ……

SubTree [S1]

x13 > 1: c (5)

x13 <= 1:

:...x18 <= 2: c (3/1)

   x18 > 2: b (3)

SubTree [S2]

x13 > 1: c (2)

x13 <= 1:

:...x13 > 0: b (4)

   x13 <= 0:

   :...x17 <= 7: b (10/2)

      x17 > 7: c (4)

Evaluation on training data (2460 cases):

```
          Decision Tree
        ----------------
        Size      Errors

         133  275(11.2%)   <<

         (a)   (b)   (c)   (d)   (e)    <-classified as
        ----  ----  ----  ----  ----

         404    56                      (a): class a
          82  1341    18          1     (b): class b
           5    62   172     2     2     (c): class c
           2    19     7    77    10     (d): class d
           1     2     4     2   191     (e): class e
```

Time: 0.1 secs

> treeModel
Call:

```
C5.0.default(x = trainset[, -28], y = trainset$y)

Classification Tree

Number of samples: 2460

Number of predictors: 27

Tree size: 137

Non-standard options: attempt to group attributes

> ruleModel <- C5.0(y ~ ., data = trainset, rules = TRUE)
> ruleModel

Call:
C5.0.formula(formula = y ~ ., data = trainset, rules = TRUE)
Rule-Based Model
Number of samples: 2460
Number of predictors: 27

Number of Rules: 63
Non-standard options: attempt to group attributes
> summary(ruleModel)
Call:
C5.0.formula(formula = y ~ ., data = trainset, rules = TRUE)
C5.0 [Release 2.07 GPL Edition]        Fri Apr 07 17:53:31 2017
-----------------------------
Class specified by attribute `outcome'
Read 2460 cases (28 attributes) from undefined.data

Rules:

Rule 1: (47, lift 5.2)
      x1 = 2
      x3 = 79
      ->  class a  [0.980]

Rule 2: (84/2, lift 5.2)
      x1 = 3
      x3 in {31, 36}
      x4 = 0
      ->  class a  [0.965]

Rule 3: (26, lift 5.2)
      x3 = 79
      x7 = 1
      x31 <= 593
      ->  class a  [0.964]

Rule 4: (42/3, lift 4.9)
      x2 in {2, 3}
      x3 = 79
      x13 <= 0
      ->  class a  [0.909]

Rule 5: (55/12, lift 4.1)
      x1 = 3
      x2 in {2, 3}
      x3 in {21, 32, 36, 45}
      x4 = 0
```

……

Evaluation on training data (2460 cases):

Rules

No Errors

63 302(12.3%) <<

(a) (b) (c) (d) (e) <-classified as

---

402 58 (a): class a

83 1336 21 2 (b): class b

11 71 156 2 3 (c): class c

2 20 6 74 13 (d): class d

1 3 4 2 190 (e): class e

```
> summary(c50pred)
  a   b   c   d   e
125 373  48  19  50
> table(predict=c50pred, truth=testset$y)
       truth
predict   a   b   c   d   e
      a  97  25   2   0   1
      b  13 325  29   5   1
      c   0  15  28   4   1
      d   0   0   1   8  10
      e   0   0   2  12  36
> acc=mean(c50pred==testset$y)
> acc
[1] 0.803252
```

SES AB veri seti için sonuçlar

```
>treeModel <- C5.0(x = trainset[, -28], y = trainset$y)

> C5imp(treeModel,metric = "usage")
   Overall                          x18   11.63
x2   100.00                         x30    8.13
x3    98.98                         x12    8.01
x1    81.99                         x16    4.39
x19   51.38                         x29    3.41
x7    46.42                         x15    2.36
x31   41.99                         x20    2.28
x13   31.79                         x23    1.95
x21   26.26                         x26    1.91
x17   22.80                         x25    1.83
x4    21.26                         x24    0.41
x27   18.21                         x11    0.28
x28   17.03                         x6     0.00
x14   16.02
x22    0.00

summary(treeModel)

Call:
C5.0.default(x = trainset[, -28], y = trainset$y)


C5.0 [Release 2.07 GPL Edition]        Fri Apr 07 17:57:53 2017
-------------------------------

Class specified by attribute `outcome'

Read 2460 cases (28 attributes) from undefined.data

Decision tree:

x2 = 1:
:...x1 in {2,3}:
:   :...x19 > 29: d (2)
:   :   x19 <= 29:
:   :   :...x19 > 6: c (2)
:   :       x19 <= 6:
:   :       :...x30 > 18: c (2/1)
:   :           x30 <= 18:
:   :           :...x31 <= 201: b (19/5)
```

```
:  :            x31 > 201:
:  :            :...x3 = 72: d (2)
:  :               x3 in {21,22,31,32,36,43,45,51,61,71,76,77,78,79,
:  :                     87}: b (2)
:  x1 = 1:
:  :...x3 in {21,22,31,32,36,43,45,51,61,71,72}: e (0)
:     x3 = 76:
:     :...x7 in {1,2,3,4,6,7}: e (211/28)
:     :  x7 = 5:
:     :  :...x17 <= 4: d (5)
:     :      x17 > 4:
:     :      :...x16 > 87: e (6)
:     :         x16 <= 87:
:     :         :...x29 > 86: e (4)
:     :            x29 <= 86:
:     :            :...x27 > 9: d (2)
:     :               x27 <= 9:
:     :               :...x30 > 1: e (3)
:     :                  x30 <= 1:
:     :                  :...x15 <= 68: d (5)
:     :                     x15 > 68: e (3/1)
:     x3 in {77,78,79,87}:
:     :...x7 in {2,7}: d (0)
:        x7 = 6:
:        :...x3 in {78,79,87}: d (4)
:        :  x3 = 77:
:        :  :...x19 <= 201: e (3)
:        :      x19 > 201: d (2)
:        x7 in {1,3,4,5}:
:        :...x3 in {79,87}: c (9)
:           x3 in {77,78}:
:           :...x31 <= 25: d (5)
:              x31 > 25:
:              :...x15 > 384: d (3)
:                 x15 <= 384:
:                 :...x30 > 43: c (6)
:                    x30 <= 43:
:                    :...x7 = 1: c (1)
:                       x7 = 5: d (1)
:                       x7 = 3:
:                       :...x15 <= 54: c (2)
:                       :  x15 > 54: d (4)
:                       x7 = 4:
:                       :...x18 <= 4: d (3)
```

: x18 > 4: c (3) …….

SubTree [S1]

x1 = 2: b (1)
x1 = 1:
:...x28 <= 282: c (4)
   x28 > 282: b (4/1)

SubTree [S2]

x26 <= 1: c (2)
x26 > 1: b (3/1)

SubTree [S3]

x12 <= 0: b (2)
x12 > 0: c (7/2)

SubTree [S4]

x25 > 0: c (2)
x25 <= 0:
:...x17 <= 0: b (6/1)
   x17 > 0: c (3/1)

SubTree [S5]

x31 > 6: b (2)
x31 <= 6:
:...x18 > 0: c (6)
   x18 <= 0:
   :...x31 <= 3: b (3)
      x31 > 3: c (3/1)

SubTree [S6]

x18 > 2: c (2)
x18 <= 2:
:...x18 > 1: b (3)
   x18 <= 1:
   :...x17 <= 0: b (2)
      x17 > 0:

```
:...x25 <= 0: c (3)
    x25 > 0: b (2)

Evaluation on training data (2460 cases):

        Decision Tree
      ----------------
      Size      Errors

       185  308(12.5%)   <<

       (a)   (b)   (c)   (d)   (e)    <-classified as
      ----  ----  ----  ----  ----
       407    49                      (a): class a
        56  1212    19     1     1    (b): class b
         7   119   249     1    13    (c): class c
         2     6    12    55    16    (d): class d
         1     2     3         229    (e): class e

>ruleModel <- C5.0(y ~ ., data = trainset, rules = TRUE)
> ruleModel

Call:
C5.0.formula(formula = y ~ ., data = trainset, rules = TRUE)

Rule-Based Model
Number of samples: 2460
Number of predictors: 27

Number of Rules: 67

Non-standard options: attempt to group attributes

> summary(ruleModel)

Call:
C5.0.formula(formula = y ~ ., data = trainset, rules = TRUE)
C5.0 [Release 2.07 GPL Edition]        Fri Apr 07 17:59:17 2017
-------------------------------
Class specified by attribute `outcome'
Read 2460 cases (28 attributes) from undefined.data
Rules:                                                  x1 = 1
                                                        x3 = 77
Rule 1: (13, lift 5.0)                                  x7 = 1
```

x14 > 0
x17 <= 14
-> class a [0.933]

Rule 2: (11, lift 5.0)
x1 = 1
x2 in {2, 3}
x3 = 77
x7 = 3
x14 > 0
x31 <= 239
-> class a [0.923]

Rule 3: (10, lift 4.9)
x3 = 77
x7 = 2
x14 > 0
x30 <= 2
-> class a [0.917]

Rule 4: (9, lift 4.9)
x1 in {1, 2}
x3 in {77, 79}
x7 = 7
x14 > 0
x19 > 0
x27 <= 0
x31 <= 105
-> class a [0.909]

Rule 5: (81/7, lift 4.9)
x3 in {31, 36}
-> class a [0.904]

Rule 6: (8, lift 4.9)
x1 = 1
x3 = 77
x7 = 7
x14 > 4
x28 <= 134
x31 <= 239
-> class a [0.900]

Rule 7: (16/1, lift 4.8)
x1 in {1, 2}
x3 in {72, 77, 79}
x27 > 2
x28 <= 134
x31 <= 239
-> class a [0.889]

Rule 8: (6, lift 4.7)
x1 = 1
x2 in {2, 3}
x3 in {72, 77}
x7 = 5
x18 > 4
-> class a [0.875]

Rule 9: (6, lift 4.7)
x1 = 3
x3 = 79
x23 <= 0
-> class a [0.875]

Rule 10: (12/1, lift 4.6)
x1 = 1
x2 in {2, 3}
x3 = 77
x7 = 4
x14 > 0
x30 <= 2
-> class a [0.857]

Rule 11: (4, lift 4.5)
x3 = 45
x17 > 5
x17 <= 38
x20 > 0
-> class a [0.833]

Rule 12: (4, lift 4.5)
x3 = 61
x7 in {1, 2, 4}
x31 > 0
-> class a [0.833]

……

```
Evaluation on training data (2460 cases):
        Rules
      ----------------
        No      Errors
        67  373(15.2%)   <<
       (a)   (b)   (c)   (d)   (e)    <-classified as
      ----  ----  ----  ----  ----
       378    78                     (a): class a
        45  1213    30          1    (b): class b
         8   141   226          14    (c): class c
         2     7    20    42    20    (d): class d
               5     2         228    (e): class e

> summary(c50pred)
  a   b   c   d   e
114 367  75   9  50
> table(predict=c50pred, truth=testset$y)
       truth
predict   a   b   c   d   e
      a  84  28   0   1   1
      b  16 303  47   0   1
      c   0  34  38   3   0
      d   0   0   2   5   2
      e   0   0   5   3  42
> acc=mean(c50pred==testset$y)
> acc
[1] 0.7674797
```

20+ABC1 veri seti için sonuçlar

```
> treeModel <- C5.0(x = trainset[, -28], y = trainset$y)
> C5imp(treeModel,metric = "usage")
   Overall                                    x26    8.33
x2   100.00                                   x19    8.25
x3    99.35                                   x12    7.80
x1    89.15                                   x18    6.75
x17   51.50                                   x21    4.15
x7    41.18                                   x20    4.07
x16   17.20                                   x15    3.21
x4    13.09                                   x28    2.93
x11   11.79                                   x25    2.52
x13   11.46                                   x22    1.99
x14   11.18                                   x29    1.34
x27   10.77                                   x24    1.10
x31   10.33                                   x23    0.45
x30    8.50                                   x6     0.00
summary(treeModel)

Call:
C5.0.default(x = trainset[, -28], y = trainset$y)
C5.0 [Release 2.07 GPL Edition]        Fri Apr 07 18:00:14 2017
-------------------------------
Class specified by attribute `outcome'
Read 2460 cases (28 attributes) from undefined.data
Decision tree:

x2 = 1:
:...x1 in {2,3}:
:   :...x1 = 2:
:   :   :...x24 <= 4: c (13/3)
:   :   :   x24 > 4: d (3/1)
:   :   x1 = 3:
:   :   :...x3 = 21: d (1)
:   :       x3 in {22,31,32,36,43,45,51,61,71,72,76,77,78,79,87}: b (13)
:   x1 = 1:
:   :...x3 in {21,22,31,32,36,43,45,51,61,71,72,76}: e (239/33)
:       x3 in {79,87}: c (8)
:       x3 in {77,78}:
:       :...x31 > 363:
:           :...x19 <= 201: e (4)
:           :   x19 > 201: d (2)
:           x31 <= 363:
:           :...x30 <= 17: d (16)
```

```
:              x30 > 17:
:              :...x3 = 78: c (4)
:                 x3 = 77:
:                 :...x21 <= 0: d (5)
:                    x21 > 0: c (3/1)
x2 in {2,3}:
:...x3 in {21,31,36,45,61,72,77,79}:
   :...x3 in {21,45,61,72,79}:
   :   :...x1 in {1,2}:
   :   :   :...x14 <= 2:
   :   :   :   :...x11 <= 7:
   :   :   :   :   :...x16 <= 30: a (188/42)
   :   :   :   :   :   x16 > 30: c (5/1)
   :   :   :   :   x11 > 7:
   :   :   :   :   :...x24 > 0: a (4)
   :   :   :   :       x24 <= 0:
   :   :   :   :       :...x15 <= 26: c (5)
   :   :   :   :           x15 > 26: d (2)
   :   :   :   x14 > 2:
   :   :   :   :...x30 > 22: b (2)
   :   :   :       x30 <= 22:
   :   :   :       :...x7 in {1,2,3,4,5}: c (0)
   :   :   :           x7 = 6: d (4/2)
   :   :   :           x7 = 7:
   :   :   :           :...x26 <= 37: c (5/1)
   :   :   :               x26 > 37: d (2)
   :   :   x1 = 3:
   :   :   :...x3 = 72: b (237/31)
   :   :       x3 = 21:
   :   :       :...x2 = 2: d (1)
   :   :       :   x2 = 3:
   :   :       :   :...x11 <= 0: a (4/1)
   :   :       :       x11 > 0: b (2)
   :   :       x3 in {45,61,79}:
   :   :       :...x7 in {6,7}: b (63/16)
   :   :           x7 = 4:
   :   :           :...x31 <= 4: b (18/3)
   :   :           :   x31 > 4:
   :   :           :   :...x15 <= 35: c (2)
   :   :           :       x15 > 35: a (2)
   :   :           x7 = 1:
   :   :           :...x30 > 1: b (13/2)
   :   :           :   x30 <= 1:
   :   :           :   :...x18 > 1: a (5)
```

```
:   :       :     x18 <= 1:
:   :       :     :...x28 <= 1: a (10/3)
:   :       :         x28 > 1: b (4) ……..
```

SubTree [S1]

```
x15 <= 0: b (3/1)
x15 > 0: c (2)
```

SubTree [S2]

```
x17 > 0: c (3/1)
x17 <= 0:
:...x11 <= 0: c (2)
   x11 > 0: b (2)
```

SubTree [S3]

```
x30 > 2: b (2)
x30 <= 2:
:...x17 > 1: b (8/1)
   x17 <= 1:
   :...x12 > 6: b (4)
      x12 <= 6:
      :...x13 > 0: c (4)
         x13 <= 0:
         :...x29 > 0: c (2)
            x29 <= 0:
            :...x15 <= 4: c (8/2)
               x15 > 4: b (3)
Evaluation on training data (2460 cases):
         Decision Tree
       ----------------
       Size      Errors
        183  272(11.1%)   <<
        (a)   (b)   (c)   (d)   (e)    <-classified as
       ----  ----  ----  ----  ----
        401    35     1          1    (a): class a
         64  1204    35     1     1    (b): class b
          8    57   299     3    17    (c): class c
          2    14     6    41    24    (d): class d
          1          2         243    (e): class e
ruleModel <- C5.0(y ~ ., data = trainset, rules = TRUE)
```

```
> ruleModel

Call:
C5.0.formula(formula = y ~ ., data = trainset, rules = TRUE)

Rule-Based Model
Number of samples: 2460
Number of predictors: 27

Number of Rules: 62
Non-standard options: attempt to group attributes

> summary(ruleModel)
Call:
C5.0.formula(formula = y ~ ., data = trainset, rules = TRUE)

C5.0 [Release 2.07 GPL Edition]        Fri Apr 07 18:01:26 2017
-------------------------------
Class specified by attribute `outcome'
Read 2460 cases (28 attributes) from undefined.data
Rules:

Rule 1: (14, lift 5.3)
      x3 in {36, 77}
      x7 = 5
      x15 > 2
      x26 <= 0
      ->  class a  [0.938]

Rule 2: (92/6, lift 5.2)
      x1 = 3
      x3 in {31, 36}
      ->  class a  [0.926]

Rule 3: (8, lift 5.1)
      x3 = 77
      x7 = 7
      x14 > 4
      x26 <= 3
      ->  class a  [0.900]

Rule 4: (7, lift 5.0)
      x2 = 3
      x3 = 77
      x7 = 6
      x12 > 1
      x17 <= 2
      ->  class a  [0.889]

Rule 5: (129/16, lift 4.9)
      x2 in {2, 3}
      x3 in {31, 36, 77}
      x7 in {1, 2, 3, 4}
      ->  class a  [0.870]

Rule 6: (5, lift 4.8)
      x1 = 2
      x2 = 2
      x3 = 79
      x14 <= 2
      x24 > 0
      ->  class a  [0.857]

…..
```

Evaluation on training data (2460 cases):

```
        Rules
      ----------------
       No      Errors

       62  336(13.7%)   <<

      (a)   (b)   (c)   (d)   (e)    <-classified as
      ----  ----  ----  ----  ----
       371    65     1     1          (a): class a
        60  1223    20     1     1    (b): class b
         9   102   252     3    18    (c): class c
         2    20     2    37    26    (d): class d
         1     4           241    (e): class e
```

```
> summary(c50pred)
 a   b   c   d   e
98 352  96  14  55
> table(predict=c50pred, truth=testset$y)
      truth
predict   a   b   c   d   e
      a  77  19   1   1   0
      b  14 298  35   5   0
      c   0  38  56   2   0
      d   0   0   8   6   0
      e   0   0   2   6  47
> acc=mean(c50pred==testset$y)
> acc
[1] 0.7869919
```

Ek 8: Destek vektör makineleri uygulama kod ve sonuçları

```
install.packages("e1071")
install.packages("openxlsx")
library(openxlsx)
library(e1071)
#dosyayi oku
setwd("C:\\Users\\...")
dat <- read.xlsx("veriseti.xlsx", sheet=1)
#sınıf degiskenini faktor yap
dat$y     <-factor(dat$y)
#veriyi test ve train diye ayir
index <- 1:nrow(dat)
testindex <- sample(index, trunc(length(index)/5))
testset <- dat[testindex,]
trainset <- dat[-testindex,]

##### polinom fonksiyon
#####parametre optimizasyonu
tobj <- tune.svm(y ~ ., data = trainset, kernel="polynomial", degree=c(1,2,3,4), cost = 10^(0:3), gamma = 10^(-3:1))
summary(tobj)

#model
svmpmodel <- svm(y  ~ ., data = trainset,kernel="polynomial", degree =3, cost = 100, gamma = 0.01, cross=10)
summary(svmpmodel)

#tahmin
svmppred <- predict(svmpmodel, testset[-56], decision.values = TRUE)
summary(svmppred)

#karşılaştırma matrisi
svmtable=table(predict=svmppred, truth=testset$y)
svmtable

#Dogruluk oranı
confusionMatrix(svmtable)
acc=mean(svmppred==testset$y)
acc
```

5+ veri seti için sonuçlar

```
> svmpmodel <- svm(y  ~ ., data = trainset,kernel="polynomial", degree =3, cost = 100, gamma = 0.01, cross=10)
> summary(svmpmodel)
Call:
svm(formula = y ~ ., data = trainset, kernel = "polynomial", degree = 3, cost = 100, gamma = 0.01, cross = 10)
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  100
     degree:  3
      gamma:  0.01
     coef.0:  0
Number of Support Vectors:  1291
 ( 609 240 114 109 219 )
Number of Classes:  5
Levels:
 a b c d e

10-fold cross-validation on training data:
Total Accuracy: 80.0813
Single Accuracies:
 80.0813 80.0813 73.98374 84.55285 79.6748 82.52033 82.11382 81.30081 76.82927 79.6748
> svmppred <- predict(svmpmodel, testset[-56], decision.values = TRUE)
> summary(svmppred)
  a   b   c   d   e
118 403  40  10  44
> svmtable=table(predict=svmppred, truth=testset$y)
> svmtable
       truth
predict   a   b   c   d   e
      a  95  22   1   0   0
      b  17 339  39   6   2
      c   0  10  26   3   1
      d   0   0   0   4   6
      e   0   0   2   9  33
> classAgreement(svmtable)
$diag
[1] 0.8081301
$kappa
[1] 0.6555897
$rand
```

[1] 0.786976
$crand
[1] 0.5702647

> confusionMatrix(svmtable)
Confusion Matrix and Statistics

```
      truth
predict   a   b   c   d   e
      a  95  22   1   0   0
      b  17 339  39   6   2
      c   0  10  26   3   1
      d   0   0   0   4   6
      e   0   0   2   9  33
```

Overall Statistics

```
               Accuracy : 0.8081
                 95% CI : (0.7747, 0.8385)
    No Information Rate : 0.6033
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6556
 Mcnemar's Test P-Value : NA
```

Statistics by Class:

```
                     Class: a Class: b Class: c Class: d Class: e
Sensitivity           0.8482   0.9137  0.38235 0.181818  0.78571
Specificity           0.9543   0.7377  0.97441 0.989882  0.98080
Pos Pred Value        0.8051   0.8412  0.65000 0.400000  0.75000
Neg Pred Value        0.9658   0.8491  0.92696 0.970248  0.98424
Prevalence            0.1821   0.6033  0.11057 0.035772  0.06829
Detection Rate        0.1545   0.5512  0.04228 0.006504  0.05366
Detection Prevalence  0.1919   0.6553  0.06504 0.016260  0.07154
Balanced Accuracy     0.9012   0.8257  0.67838 0.585850  0.88326
```

> acc=mean(svmppred==testset$y)
> acc
[1] 0.8081301

SES AB veri seti için sonuçlar

> svmpmodel <- svm(y ~ ., data = trainset,kernel="polynomial", degree =3, cost = 100, gamma = 0.01, cross=10)

```
> summary(svmpmodel)

Call:
svm(formula = y ~ ., data = trainset, kernel = "polynomial", degree = 3, cost = 100, gamma
= 0.01, cross = 10)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  100
     degree:  3
      gamma:  0.01
     coef.0:  0

Number of Support Vectors:  1441

 ( 365 671 234 98 73 )


Number of Classes:  5

Levels:
 a b c d e

10-fold cross-validation on training data:

Total Accuracy: 76.38211
Single Accuracies:
 75.60976 76.01626 76.82927 80.0813 76.42276 76.01626 77.64228 72.76423 73.57724
78.86179



>
> #tahmin
> svmppred <- predict(svmpmodel, testset[-56], decision.values = TRUE)
> summary(svmppred)
  a   b   c   d   e
116 369  52  15  63
>
>
> #confusion matrix
> svmtable=table(predict=svmppred, truth=testset$y)
```

```
> svmtable
      truth
predict   a   b   c   d   e
      a  89  24   2   1   0
      b  15 306  46   2   0
      c   4   8  33   2   5
      d   0   1   1  12   1
      e   1   0   2   7  53
> classAgreement(svmtable)
$diag
[1] 0.801626

$kappa
[1] 0.6766575

$rand
[1] 0.7959588

$crand
[1] 0.5715478

> confusionMatrix(svmtable)
Confusion Matrix and Statistics

      truth
predict   a   b   c   d   e
      a  89  24   2   1   0
      b  15 306  46   2   0
      c   4   8  33   2   5
      d   0   1   1  12   1
      e   1   0   2   7  53

Overall Statistics

               Accuracy : 0.8016
                 95% CI : (0.7679, 0.8324)
    No Information Rate : 0.5512
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6767
 Mcnemar's Test P-Value : NA

Statistics by Class:
```

	Class: a	Class: b	Class: c	Class: d	Class: e
Sensitivity	0.8165	0.9027	0.39286	0.50000	0.89831
Specificity	0.9466	0.7717	0.96422	0.99492	0.98201
Pos Pred Value	0.7672	0.8293	0.63462	0.80000	0.84127
Neg Pred Value	0.9599	0.8659	0.90941	0.98000	0.98913
Prevalence	0.1772	0.5512	0.13659	0.03902	0.09593
Detection Rate	0.1447	0.4976	0.05366	0.01951	0.08618
Detection Prevalence	0.1886	0.6000	0.08455	0.02439	0.10244
Balanced Accuracy	0.8816	0.8372	0.67854	0.74746	0.94016

```
> acc=mean(svmppred==testset$y)
> acc
[1] 0.801626
```

20+ABC1 veri seti için sonuçlar

```
svmpmodel <- svm(y  ~ ., data = trainset,kernel="polynomial", degree =3, cost = 100, gamma = 0.01, cross=10)
> summary(svmpmodel)

Call:
svm(formula = y ~ ., data = trainset, kernel = "polynomial", degree = 3, cost = 100, gamma = 0.01, cross = 10)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  100
     degree:  3
      gamma:  0.01
     coef.0:  0

Number of Support Vectors:  1362

 ( 338 614 98 234 78 )


Number of Classes:  5

Levels:
 a b c d e

10-fold cross-validation on training data:
```

Total Accuracy: 77.39837
Single Accuracies:
77.23577 80.89431 75.20325 73.98374 77.23577 78.45528 80.89431 78.04878 72.35772 79.6748

```
> svmppred <- predict(svmpmodel, testset[-56], decision.values = TRUE)
> summary(svmppred)
  a   b   c   d   e
117 368  58  18  54
> svmtable=table(predict=svmppred, truth=testset$y)
> svmtable
       truth
predict   a   b   c   d   e
      a  87  29   1   0   0
      b  10 296  56   6   0
      c   0  12  39   1   6
      d   1   1   0  12   4
      e   0   0   2   4  48
> classAgreement(svmtable)
$diag
[1] 0.7837398

$kappa
[1] 0.6491668

$rand
[1] 0.7720187

$crand
[1] 0.5204258

> confusionMatrix(svmtable)
Confusion Matrix and Statistics

       truth
predict   a   b   c   d   e
      a  87  29   1   0   0
      b  10 296  56   6   0
      c   0  12  39   1   6
      d   1   1   0  12   4
      e   0   0   2   4  48
```

```
Overall Statistics

               Accuracy : 0.7837
                 95% CI : (0.7491, 0.8157)
    No Information Rate : 0.5496
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6492
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: a Class: b Class: c Class: d Class: e
Sensitivity            0.8878   0.8757  0.39796  0.52174  0.82759
Specificity            0.9420   0.7401  0.96325  0.98986  0.98923
Pos Pred Value         0.7436   0.8043  0.67241  0.66667  0.88889
Neg Pred Value         0.9779   0.8300  0.89408  0.98157  0.98217
Prevalence             0.1593   0.5496  0.15935  0.03740  0.09431
Detection Rate         0.1415   0.4813  0.06341  0.01951  0.07805
Detection Prevalence   0.1902   0.5984  0.09431  0.02927  0.08780
Balanced Accuracy      0.9149   0.8079  0.68060  0.75580  0.90841
> acc=mean(svmppred==testset$y)
> acc
[1] 0.7837398

###### radyal temelli fonksiyon
#####parametre optimizasyonu
tobj <- tune.svm(y ~ ., data = trainset, kernel="radial",  gcost = 10^(0:3), gamma = 10^(-
3:1))
#summary(tobj)
#10cross
svmrmodel <- svm(y ~ ., data = trainset,kernel="radial",cost = 10, gamma = 0.01,
cross=10)
summary(svmrmodel)

#tahmin
svmrpred <- predict(svmrmodel, testset[-56])
summary(svmrpred)

#karşılaştırma matrisi
svmr=table(predict=svmrpred, truth=testset$y)

#doğruluk oranı
confusionMatrix(svmtable)
```

```
acc=mean(svmrpred==testset$y)
acc
```

5+veri seti için sonuçlar

```
> svmrmodel <- svm(y ~ ., data = trainset,kernel="radial",cost = 10, gamma = 0.01, cross=10)
> summary(svmrmodel)

Call:
svm(formula = y ~ ., data = trainset, kernel = "radial", cost = 10, gamma = 0.01, cross = 10)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  10
      gamma:  0.01

Number of Support Vectors:  1172

 ( 474 238 132 110 218 )


Number of Classes:  5

Levels:
 a b c d e

10-fold cross-validation on training data:

Total Accuracy: 80.81301
Single Accuracies:
 81.30081 85.36585 75.20325 80.4878 80.0813 78.86179 86.17886 82.11382 78.45528 80.0813
> svmrpred <- predict(svmrmodel, testset[-56])
> summary(svmrpred)
  a   b   c   d   e
124 393  36   7  55
> svmr=table(predict=svmrpred, truth=testset$y)
> classAgreement(svmr)
$diag
```

```
[1] 0.8260163

$kappa
[1] 0.6919559

$rand
[1] 0.7996187

$crand
[1] 0.5936202

> confusionMatrix(svmr)
Confusion Matrix and Statistics

          truth
predict   a   b   c   d   e
        a  97  26   1   0   0
        b  14 337  36   6   0
        c   0   8  27   1   0
        d   0   0   0   6   1
        e   1   0   4   9  41

Overall Statistics

               Accuracy : 0.826
                 95% CI : (0.7937, 0.8552)
    No Information Rate : 0.6033
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.692
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: a Class: b Class: c Class: d Class: e
Sensitivity           0.8661   0.9084  0.39706 0.272727  0.97619
Specificity           0.9463   0.7705  0.98355 0.998314  0.97557
Pos Pred Value        0.7823   0.8575  0.75000 0.857143  0.74545
Neg Pred Value        0.9695   0.8468  0.92919 0.973684  0.99821
Prevalence            0.1821   0.6033  0.11057 0.035772  0.06829
Detection Rate        0.1577   0.5480  0.04390 0.009756  0.06667
Detection Prevalence  0.2016   0.6390  0.05854 0.011382  0.08943
Balanced Accuracy     0.9062   0.8394  0.69030 0.635520  0.97588
> acc=mean(svmrpred==testset$y)
```

> acc
[1] 0.8260163

SES AB veri seti için sonuçlar

> svmrmodel <- svm(y ~ ., data = trainset,kernel="radial",cost = 10, gamma = 0.01, cross=10)
> summary(svmrmodel)

Call:
svm(formula = y ~ ., data = trainset, kernel = "radial", cost = 10, gamma = 0.01, cross = 10)

Parameters:
  SVM-Type: C-classification
 SVM-Kernel: radial
       cost: 10
      gamma: 0.01

Number of Support Vectors: 1392

 (369 594 238 115 76)

Number of Classes: 5

Levels:
 a b c d e

10-fold cross-validation on training data:

Total Accuracy: 77.15447
Single Accuracies:
 76.82927 76.42276 78.86179 75.20325 78.86179 75.20325 76.82927 78.45528 74.79675 80.0813

> svmrpred <- predict(svmrmodel, testset[-56])
> summary(svmrpred)
  a   b   c   d   e
111 378 43 17 66
> svmr=table(predict=svmrpred, truth=testset$y)
> classAgreement(svmr)
$diag

```
[1] 0.803252

$kappa
[1] 0.676555

$rand
[1] 0.8003337

$crand
[1] 0.5840057

> confusionMatrix(svmr)
Confusion Matrix and Statistics

          truth
predict   a   b   c   d   e
        a  89  19   2   1   0
        b  19 312  46   1   0
        c   1   7  30   2   3
        d   0   1   3  10   3
        e   0   0   3  10  53

Overall Statistics

               Accuracy : 0.8033
                 95% CI : (0.7696, 0.834)
    No Information Rate : 0.5512
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6766
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: a Class: b Class: c Class: d Class: e
Sensitivity           0.8165   0.9204  0.35714  0.41667  0.89831
Specificity           0.9565   0.7609  0.97552  0.98816  0.97662
Pos Pred Value        0.8018   0.8254  0.69767  0.58824  0.80303
Neg Pred Value        0.9603   0.8861  0.90559  0.97659  0.98907
Prevalence            0.1772   0.5512  0.13659  0.03902  0.09593
Detection Rate        0.1447   0.5073  0.04878  0.01626  0.08618
Detection Prevalence  0.1805   0.6146  0.06992  0.02764  0.10732
Balanced Accuracy     0.8865   0.8406  0.66633  0.70241  0.93746
> acc=mean(svmrpred==testset$y)
```

> acc
[1] 0.803252

20+ABC1 veri seti için sonuçlar

> svmrmodel <- svm(y ~ ., data = trainset,kernel="radial",cost = 10, gamma = 0.01, cross=10)
> summary(svmrmodel)

Call:
svm(formula = y ~ ., data = trainset, kernel = "radial", cost = 10, gamma = 0.01, cross = 10)

Parameters:
  SVM-Type: C-classification
 SVM-Kernel: radial
       cost: 10
      gamma: 0.01

Number of Support Vectors: 1302

 (347 534 110 229 82)

Number of Classes: 5

Levels:
 a b c d e

10-fold cross-validation on training data:

Total Accuracy: 78.41463
Single Accuracies:
 79.26829 77.23577 78.04878 75.20325 76.42276 84.14634 82.92683 76.82927 79.26829 74.79675

> svmrpred <- predict(svmrmodel, testset[-56])
> summary(svmrpred)
  a   b   c   d   e
112 376 51 12 64
> svmr=table(predict=svmrpred, truth=testset$y)
> classAgreement(svmr)
$diag

```
[1] 0.798374

$kappa
[1] 0.67012

$rand
[1] 0.7789148

$crand
[1] 0.5385086

> confusionMatrix(svmr)
Confusion Matrix and Statistics

          truth
predict   a   b   c   d   e
        a 87  25   0   0   0
        b 11 301  60   4   0
        c  0  12  35   3   1
        d  0   0   1  11   0
        e  0   0   2   5  57

Overall Statistics

               Accuracy : 0.7984
                 95% CI : (0.7645, 0.8294)
    No Information Rate : 0.5496
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6701
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: a Class: b Class: c Class: d Class: e
Sensitivity            0.8878   0.8905  0.35714  0.47826  0.98276
Specificity            0.9516   0.7292  0.96905  0.99831  0.98743
Pos Pred Value         0.7768   0.8005  0.68627  0.91667  0.89062
Neg Pred Value         0.9781   0.8452  0.88830  0.98010  0.99819
Prevalence             0.1593   0.5496  0.15935  0.03740  0.09431
Detection Rate         0.1415   0.4894  0.05691  0.01789  0.09268
Detection Prevalence   0.1821   0.6114  0.08293  0.01951  0.10407
Balanced Accuracy      0.9197   0.8099  0.66310  0.73829  0.98510
> acc=mean(svmrpred==testset$y)
```

```
> acc
[1] 0.798374
```

#####**lineer fonksiyon**

```
svmpmodel <- svm(y ~ ., data = trainset, kernel="linear", cross=10)
summary(svmpmodel)

#tahmin
svmppred <- predict(svmpmodel, testset[-56])
summary(svmppred)

#karşılaştırma matrisi
svml=table(predict=svmppred, truth=testset$y)

#doğruluk oranı
confusionMatrix(svmtable)
acc=mean(svmppred==testset$y)
acc
```

5+ veri seti için sonuçlar

```
> svmpmodel <- svm(y ~ ., data = trainset, kernel="linear", cross=10)
> summary(svmpmodel)

Call:
svm(formula = y ~ ., data = trainset, kernel = "linear", cross = 10)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear

Number of Support Vectors:  1070

 ( 436 225 96 100 213 )

Number of Classes:  5

Levels:
 a b c d e

10-fold cross-validation on training data:

Total Accuracy: 80.73171
Single Accuracies:
```

80.0813 74.79675 85.77236 82.92683 79.26829 78.04878 78.86179 84.55285 82.92683
80.0813

```
> svmppred <- predict(svmpmodel, testset[-56])
> summary(svmppred)
  a   b   c   d   e
129 389  35   9  53
> svml=table(predict=svmppred, truth=testset$y)
> classAgreement(svml)
$diag
[1] 0.8160163

$kappa
[1] 0.6934369

$rand
[1] 0.8027701

$crand
[1] 0.5992299

> confusionMatrix(svml)
Confusion Matrix and Statistics

       truth
predict   a   b   c   d   e
      a  98  30   1   0   0
      b  14 336  34   5   0
      c   0   5  28   2   0
      d   0   0   1   6   2
      e   0   0   4   9  40

Overall Statistics

               Accuracy : 0.816
                 95% CI : (0.7937, 0.8552)
    No Information Rate : 0.6033
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6934
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: a Class: b Class: c Class: d Class: e
```

```
Sensitivity          0.8750   0.9057  0.41176 0.272727  0.95238
Specificity          0.9384   0.7828  0.98720 0.994941  0.97731
Pos Pred Value       0.7597   0.8638  0.80000 0.666667  0.75472
Neg Pred Value       0.9712   0.8451  0.93103 0.973597  0.99644
Prevalence           0.1821   0.6033  0.11057 0.035772  0.06829
Detection Rate       0.1593   0.5463  0.04553 0.009756  0.06504
Detection Prevalence   0.2098   0.6325  0.05691 0.014634  0.08618
Balanced Accuracy      0.9067   0.8442  0.69948 0.633834  0.96485
> acc=mean(svmppred==testset$y)
> acc
[1] 0.8160163
```

SES AB veri seti için sonuçlar

```
> svmpmodel <- svm(y ~ ., data = trainset, kernel="linear", cross=10)
> summary(svmpmodel)

Call:
svm(formula = y ~ ., data = trainset, kernel = "linear", cross = 10)

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
Number of Support Vectors:  1240

 ( 379 501 214 73 73 )

Number of Classes:  5

Levels:
 a b c d e

10-fold cross-validation on training data:

Total Accuracy: 74.87805
Single Accuracies:
 77.23577 73.98374 78.86179 74.39024 71.95122 75.60976 78.86179 76.42276 69.9187
71.54472



> svmppred <- predict(svmpmodel, testset[-56])
> summary(svmppred)
  a   b   c   d   e
```

```
111 405  15  13  71
> svml=table(predict=svmppred, truth=testset$y)
> classAgreement(svml)
$diag
[1] 0.7642276

$kappa
[1] 0.6002376

$rand
[1] 0.7644395

$crand
[1] 0.5236211

> confusionMatrix(svml)
Confusion Matrix and Statistics

      truth
predict   a   b   c   d   e
     a  87  20   3   1   0
     b  20 315  66   3   1
     c   2   4   6   3   0
     d   0   0   5   6   2
     e   0   0   4  11  56

Overall Statistics

               Accuracy : 0.7642
                 95% CI : (0.7286, 0.7973)
    No Information Rate : 0.5512
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6002
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: a Class: b Class: c Class: d Class: e
Sensitivity           0.7982   0.9292 0.071429 0.250000  0.94915
Specificity           0.9526   0.6739 0.983051 0.988156  0.97302
Pos Pred Value        0.7838   0.7778 0.400000 0.461538  0.78873
Neg Pred Value        0.9563   0.8857 0.870000 0.970100  0.99449
Prevalence            0.1772   0.5512 0.136585 0.039024  0.09593
```

```
Detection Rate         0.1415   0.5122 0.009756 0.009756  0.09106
Detection Prevalence   0.1805   0.6585 0.024390 0.021138  0.11545
Balanced Accuracy      0.8754   0.8016 0.527240 0.619078  0.96109
> acc=mean(svmppred==testset$y)
> acc
[1] 0.7642276
```

20+ABC1 veri seti için sonuçlar

```
> svmpmodel <- svm(y ~ ., data = trainset, kernel="linear", cross=10)
> summary(svmpmodel)

Call:
svm(formula = y ~ ., data = trainset, kernel = "linear", cross = 10)

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear

Number of Support Vectors:  1193

 ( 354 488 66 206 79 )

Number of Classes:  5

Levels:
 a b c d e

10-fold cross-validation on training data:

Total Accuracy: 76.82927
Single Accuracies:
 75.20325 84.14634 77.23577 74.79675 76.82927 78.04878 79.6748 76.42276 71.54472
74.39024

> svmppred <- predict(svmpmodel, testset[-56])
> summary(svmppred)
  a   b   c   d   e
110 402  27  10  66
> svml=table(predict=svmppred, truth=testset$y)
> classAgreement(svml)
$diag
[1] 0.7853659
$kappa
```

```
[1] 0.6389834
$rand
[1] 0.7671884
$crand
[1] 0.5271016
> confusionMatrix(svml)
Confusion Matrix and Statistics

          truth
predict   a   b   c   d   e
        a  84  21   5   0   0
        b  14 312  72   4   0
        c   0   5  19   3   0
        d   0   0   0  10   0
        e   0   0   2   6  58

Overall Statistics

               Accuracy : 0.7854
                 95% CI : (0.7508, 0.8172)
    No Information Rate : 0.5496
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.639
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: a Class: b Class: c Class: d Class: e
Sensitivity           0.8571   0.9231  0.19388  0.43478  1.00000
Specificity           0.9497   0.6751  0.98453  1.00000  0.98564
Pos Pred Value        0.7636   0.7761  0.70370  1.00000  0.87879
Neg Pred Value        0.9723   0.8779  0.86565  0.97851  1.00000
Prevalence            0.1593   0.5496  0.15935  0.03740  0.09431
Detection Rate        0.1366   0.5073  0.03089  0.01626  0.09431
Detection Prevalence  0.1789   0.6537  0.04390  0.01626  0.10732
Balanced Accuracy     0.9034   0.7991  0.58920  0.71739  0.99282
> acc=mean(svmppred==testset$y)
> acc
[1] 0.7853659
```

Ek 9: Yapay sinir ağları uygulama kod ve sonuçları

```
library(openxlsx)
library(RSNNS)
setwd("C:\\Users\\....")
dattt <- read.xlsx("veriseti.xlsx", sheet=1)
#veriyi test ve train diye ayir
dattt <- dattt[sample(1:nrow(dattt),length(1:nrow(dattt))),1:ncol(dattt)]
datValues <- dattt[,1:55]
datTargets <- decodeClassLabels(dattt[,56])
dattt <- splitForTrainingAndTest(datValues, datTargets, ratio=0.2)
dattt <- normTrainingAndTestSet(dattt)
#model
model <- mlp(dattt$inputsTrain, dattt$targetsTrain, size=10, learnFunc="BackpropMomentum",
        learnFuncParams=c(0.01, 0.3),
        maxit=100, inputsTest=dattt$inputsTest, targetsTest=dattt$targetsTest)
summary(model)
weightMatrix(model)
#tahmin
annpred <- predict(model,dattt$inputsTest,type(class))
mattt<-confusionMatrix(dattt$targetsTest, annpred)
mattt
accuracy <- sum(diag(mattt)) / sum(mattt)
accuracy
```

5+ veri seti için sonuçlar

,"x"

1,"SNNS network definition file V1.4-3D

generated at Fri Apr 07 17:01:59 2017

network name : RSNNS_untitled

source files :

no. of units : 70

no. of connections : 600

no. of unit types : 0

no. of site types : 0

> model <- mlp(dattt$inputsTrain, dattt$targetsTrain, size=10, learnFunc="BackpropMomentum",

+ learnFuncParams=c(0.01, 0.3),

+ maxit=100, inputsTest=dattt$inputsTest, targetsTest=dattt$targetsTest)

> annpred <- predict(model,dattt$inputsTest,type(class))

> mattt<-confusionMatrix(dattt$targetsTest, annpred)

> mattt

predictions

targets 1 2 3 4 5

1 113 15 0 0 0

2 19 330 4 0 0

3 2 34 21 1 4

4 1 3 2 8 10

5 0 0 1 3 44

> accuracy <- sum(diag(mattt)) / sum(mattt)

> accuracy

[1] 0.8390244

SES AB veri seti için sonuçlar

,"x"

1,"SNNS network definition file V1.4-3D

generated at Fri Apr 07 17:04:04 2017

network name : RSNNS_untitled

source files :

no. of units : 70

no. of connections : 600

no. of unit types : 0

no. of site types : 0

> model <- mlp(dattt$inputsTrain, dattt$targetsTrain, size=10, learnFunc="BackpropMomentum",

+ learnFuncParams=c(0.01, 0.3),

+ maxit=100, inputsTest=dattt$inputsTest, targetsTest=dattt$targetsTest)

> annpred <- predict(model,dattt$inputsTest,type(class))

> mattt<-confusionMatrix(dattt$targetsTest, annpred)

> mattt

predictions

targets 1 2 3 4 5

1 96 20 1 0 0

2 20 316 9 0 0

3 0 46 29 1 4

4 0 2 3 4 11

5 0 1 0 0 52

> accuracy <- sum(diag(mattt)) / sum(mattt)

> accuracy

[1] 0.8081301

20+ABC1 veri seti için sonuçlar

1,"SNNS network definition file V1.4-3D

generated at Fri Apr 07 17:00:20 2017

network name : RSNNS_untitled

source files :

no. of units : 70

no. of connections : 600

no. of unit types : 0

no. of site types : 0

model <- mlp(dattt$inputsTrain, dattt$targetsTrain, size=10, learnFunc="BackpropMomentum",

+ learnFuncParams=c(0.01, 0.3),

+ maxit=100, inputsTest=dattt$inputsTest, targetsTest=dattt$targetsTest)

> #extractNetInfo(model)

> #par(mfrow=c(2,2))#plotIterativeError(model)

> annpred <- predict(model,dattt$inputsTest,type(class))

> mattt<-confusionMatrix(dattt$targetsTest, annpred)

> mattt

predictions

targets 1 2 3 4 5

1 89 19 0 0 0

2 20 294 16 0 0

3 0 42 48 0 4

4 1 2 3 6 8

5 1 0 0 1 61

> accuracy <- sum(diag(mattt)) / sum(mattt)

> accuracy

[1] 0.8097561

Ek 10: Yapay sinir ağları ağırlık matrisi

5+

	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10
x1	-0.859	-0.633	-0.999	-0.389	-0.03	0.005	0.604	-0.424	0.268	-0.266
x2	0.104	-0.234	0.012	0.556	0.017	-0.265	-0.668	0.556	-0.336	-0.052
x3	0.682	0.627	0.678	-0.217	-0.117	0.419	0.157	-0.008	0.012	0.181
x4	0.155	-0.79	-2.282	-0.85	-0.003	0.486	0.402	0.021	0.245	0.257
x5	0.14	0.144	0.712	0.123	0.24	0.001	-0.283	0.158	0.173	0.003
x6	0.208	0.47	0.376	0.056	-0.185	-0.058	0.046	-0.165	-0.235	-0.067
x7	0.151	-0.114	-0.098	0.155	0.31	-0.266	0.288	-0.487	0.066	-0.089
x8	1.374	0.786	0.467	0.411	0.436	0.116	-0.75	-0.68	-0.073	0.43
x9	-0.494	-0.467	0.038	0.337	0.375	-0.27	0.474	-0.509	-0.181	-0.28
x10	0.151	-0.036	0.183	-0.118	0.231	0.123	0.25	-0.326	0.284	-0.184
x11	-0.637	-0.678	0.132	0.11	0.397	-0.304	0.508	-0.878	-0.181	-0.372
x12	0.567	0.231	-0.068	-0.645	-0.198	0.691	-0.333	-0.126	-0.014	0.442
x13	-0.476	0	0.309	0.762	0.32	-0.133	0.239	-0.439	-0.505	-0.021
x14	0.632	0.461	0.384	-0.614	-0.41	0.697	-0.148	0.004	0.072	0.569
x15	0.376	0.14	0.138	0.24	0.208	-0.334	-0.001	-0.061	0.074	-0.102
x16	-0.269	0.831	-0.171	-0.364	-0.749	-0.058	-0.216	0.938	-0.27	0.448
x17	-0.904	-0.319	0.843	1.275	0.004	-0.365	0.216	0.181	-0.605	-0.159
x18	-0.056	-0.925	-1.443	-0.909	-0.238	0.947	-0.522	-0.239	0.889	-0.074
x19	0.9	0.254	1.456	-0.087	0.195	-0.228	-0.257	-0.257	-0.111	-0.322
x20	0.444	-0.472	-0.34	-0.257	0.817	-0.54	-0.049	-0.456	0.342	-0.261
x21	-1.201	-0.776	-0.239	0.335	0.33	-0.995	1.201	1.22	0.135	-1.139
x22	-0.153	-0.058	-0.327	-0.033	-0.364	-0.25	0.219	0.311	0.315	-0.137
x23	-0.137	-0.515	-0.624	0.543	0.212	-0.367	0.18	0.403	0.019	-0.646
x24	0.623	0.63	0.458	-0.496	-0.231	0.657	0.399	-0.979	-0.03	0.483
x25	-0.565	-0.52	-0.795	-0.307	0.145	-0.319	0.385	-0.499	0.277	-0.132
x26	0.046	-0.077	-0.072	0.367	0.117	-0.448	-0.391	0.652	0.147	-0.32
x27	1.02	0.731	0.833	-0.235	-0.023	0.926	0.255	0.236	-0.285	0.6
x28	-0.216	0.416	-0.005	0.102	0.132	0.174	0.051	0.349	-0.093	0.244
x29	-0.442	0.01	0.333	-0.143	-0.211	-0.08	0.175	-0.476	0.003	0.291
x30	0.122	0.095	0.368	-0.027	-0.157	0.07	0.518	0.121	0.055	-0.162
x31	-0.393	-0.207	0.097	-0.017	-0.389	0.065	0.439	0.653	-0.144	0.106
x32	-0.599	-0.339	-0.886	0.077	0.905	-0.422	-0.979	-0.389	-0.371	0.363
x33	0.177	0.023	-0.298	-0.222	0.21	0.13	-0.441	-0.136	0.278	-0.249
x34	0.475	0.087	-0.25	-0.442	-0.246	0.128	0.273	0.354	-0.012	-0.226
x35	-0.264	-0.025	-0.357	0.05	0.263	0.388	0.183	0.161	0.305	0.261
x36	-0.032	-0.069	-0.098	-0.057	-0.064	-0.113	0.021	-0.168	-0.088	0.009

x37	0.084	-0.321	-0.429	-0.331	0.692	0.116	0.057	0.692	0.255	0.276
x38	-0.47	-0.184	-0.04	0.021	0.27	0.068	-0.056	0.143	0.279	0.049
x39	-0.125	0.353	-0.356	-0.667	-0.437	-0.276	0.156	0.053	0.19	-0.899
x40	0.252	0.081	-0.246	-0.007	-0.26	-0.303	0.153	-0.102	0.291	-0.655
x41	0.265	-0.193	-0.024	-0.125	-0.582	-0.176	0.47	-0.367	0.086	0.07
x42	0.245	-0.114	0.201	0.406	-0.836	0.308	-0.125	0.173	-0.457	0.123
x43	-0.115	-0.175	-0.026	0.15	-0.194	-0.172	0.224	-0.296	0.137	0.183
x44	-0.271	-0.198	0.107	-0.303	-0.071	0.201	-0.088	-0.214	-0.058	0.009
x45	-0.283	-0.5	-0.023	0.082	-0.239	-0.044	0.156	-0.32	-0.125	0.017
x46	0.156	-0.245	-0.049	0.075	-0.555	0.035	0.064	-0.4	-0.221	-0.093
x47	-0.222	-0.316	0.477	-0.384	0.084	0.156	-0.007	-0.073	0.133	0.174
x48	-0.286	-0.183	-0.355	0.292	-0.087	0.371	0.355	-0.13	-0.144	0.367
x49	0.091	-0.152	-0.251	0.287	-0.373	0.375	-0.294	0.299	0.29	-0.384
x50	0.083	0.099	-0.441	-0.153	-0.298	-0.163	0.01	-0.294	0.582	-0.404
x51	-0.231	0.107	0.302	0.125	0.231	-0.008	0.135	-0.251	-0.032	-0.238
x52	-0.088	-0.035	0.075	-0.289	-0.25	-0.043	-0.06	-0.291	0.58	-0.747
x53	-0.285	-0.197	0.107	-0.031	-0.598	-0.087	-0.034	0.047	0.645	-0.466
x54	0.195	-0.098	-0.112	-0.153	0.028	0.111	0.21	0.131	0.067	-0.381
x55	-0.453	-0.525	-0.072	-0.337	-0.179	0.767	0.414	-0.311	0.403	-0.126

	Output_a	Output_b	Output_c	Output_d	Output_e
h1	-1.827	1.608	-1.031	0.508	-1.179
h2	-0.834	1.879	-0.862	-1.292	-1.366
h3	2.25	1.553	-3.351	-1.982	-1.504
h4	2.008	-1.649	0.556	-1.693	-1.52
h5	0.816	-1.591	-0.448	1.579	-1.655
h6	-2.231	-0.08	-1.436	-0.912	2.11
h7	0.855	-2.334	-0.532	-0.111	0.794
h8	-1.45	0.949	2.679	-1.786	-2.039
h9	-1.875	-1.228	0.374	0.507	0.607
h10	-1.819	1.215	-0.701	-0.293	-0.699

SES AB

	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10
x1	0.582	-0.073	0.722	0.087	0.075	-0.473	-1.382	-0.246	-0.187	-0.786
x2	-0.189	-0.285	0.016	0.070	-0.164	0.474	0.036	0.083	0.587	0.186
x3	-0.016	0.576	-0.734	0.627	0.345	-0.049	0.954	0.622	-0.128	0.318
x4	-0.032	-0.731	0.519	0.046	-0.424	-0.642	-0.222	0.220	-0.876	-0.827
x5	-0.475	0.048	-0.233	-0.059	0.065	-0.158	0.593	-0.231	0.793	0.258

x6	-0.033	0.352	-0.259	-0.292	-0.259	0.379	-0.096	0.028	0.076	0.088
x7	0.066	-0.301	0.229	-0.201	0.218	0.297	0.143	-0.248	0.413	-0.137
x8	-0.162	0.910	-0.445	0.336	0.297	-1.064	0.826	0.674	0.138	0.098
x9	0.618	-0.485	0.382	-0.469	0.270	0.233	-0.806	-0.686	0.188	0.187
x10	-0.189	-0.432	-0.088	-0.082	-0.317	0.217	-0.281	0.166	0.207	0.156
x11	0.544	-0.292	0.693	-0.671	0.298	0.435	-0.795	-0.914	0.332	0.170
x12	-0.102	0.733	-0.286	0.400	-0.113	-0.447	0.525	0.466	-0.544	0.252
x13	0.314	0.234	0.237	-0.398	0.108	0.352	-0.360	0.231	0.332	0.369
x14	0.424	0.473	-0.652	0.339	0.336	-0.244	0.020	0.071	-0.738	0.109
x15	-0.175	-0.788	0.584	-0.314	-0.339	-0.324	0.157	-0.862	-0.020	0.180
x16	-0.925	0.233	-0.752	0.395	0.168	0.031	0.745	0.840	0.216	-0.358
x17	0.127	0.227	-0.125	-0.449	0.477	0.795	-0.780	0.102	0.326	1.508
x18	-0.086	-1.186	0.511	0.664	-0.283	-0.877	0.110	-0.078	-0.364	-1.678
x19	0.061	0.716	-0.461	0.229	0.156	0.096	-0.345	-0.575	-0.319	0.777
x20	-0.564	-0.504	0.274	-0.600	-0.473	0.275	0.083	0.552	-0.187	0.579
x21	1.182	0.395	0.457	-0.693	-0.563	1.149	-1.100	-0.586	0.329	-0.163
x22	-0.267	-0.404	0.013	0.217	-0.373	0.278	-0.249	0.314	0.367	-0.391
x23	-0.483	0.564	0.652	-0.453	0.130	0.580	-0.558	-0.069	0.146	0.048
x24	-0.045	-0.325	-0.401	0.381	0.146	-0.530	0.524	-0.325	-0.211	-0.214
x25	0.091	0.008	0.790	-0.111	0.058	-0.466	-1.049	-0.601	-0.353	-0.773
x26	-0.503	-0.299	-0.040	0.059	-0.417	0.119	0.511	0.071	0.040	0.378
x27	0.010	0.237	-0.440	0.421	0.249	-0.089	0.799	0.449	-0.410	0.649
x28	0.292	-0.162	-0.288	-0.087	0.722	0.008	-0.688	-0.026	0.442	-0.049
x29	0.458	0.084	-0.171	0.034	0.221	-0.330	-0.523	0.026	-0.017	-0.071
x30	0.381	0.963	0.080	-0.108	0.234	0.086	0.250	0.260	-0.092	0.549
x31	-0.624	0.254	0.566	-0.335	0.122	-0.166	-0.021	0.265	-0.422	0.023
x32	-0.859	-0.103	1.086	-0.377	0.343	0.222	0.893	-0.082	-0.327	-0.456
x33	0.131	-0.988	-0.551	0.381	-0.350	-0.501	0.312	-0.162	-0.175	-0.120
x34	0.554	-0.679	-0.339	0.332	-0.858	0.504	0.081	-0.457	-0.225	0.011
x35	0.021	-0.203	0.160	0.581	-0.031	-0.182	0.063	0.321	-0.061	-0.372
x36	0.553	-0.212	0.226	0.100	-0.150	0.030	0.190	0.003	-0.431	0.017
x37	0.461	0.119	0.237	-0.424	-0.303	0.037	0.006	0.188	-0.284	-0.183
x38	0.171	-0.111	-0.111	-0.519	0.089	-0.023	0.166	0.467	-0.141	-0.289
x39	0.028	-0.162	-0.306	-0.324	-0.217	0.035	-0.276	0.473	-0.269	-0.267
x40	0.084	-0.438	0.164	0.166	-0.193	-0.141	-0.073	0.289	0.200	-0.463
x41	-0.235	-0.211	0.332	0.146	-0.338	-0.005	-0.255	0.366	-0.324	-0.293
x42	0.017	0.338	-0.548	0.073	-0.119	0.175	-0.402	0.663	-0.313	0.137
x43	0.184	-0.140	0.036	0.099	0.089	-0.234	-0.236	-0.033	-0.053	0.018
x44	-0.074	-0.088	0.090	0.102	0.021	-0.088	-0.202	-0.241	-0.028	-0.149
x45	-0.225	-0.177	0.182	0.064	0.111	-0.065	-0.134	-0.112	-0.455	-0.029

x46	0.239	-0.017	0.331	-0.305	-0.191	-0.261	-0.247	-0.183	-0.246	-0.345
x47	0.075	0.125	0.028	0.010	-0.075	0.034	0.343	-0.173	-0.044	-0.415
x48	0.186	-0.128	-0.215	0.251	0.298	-0.179	0.182	0.138	-0.140	0.064
x49	-0.385	-0.285	0.097	-0.025	-0.038	0.319	0.493	-0.259	0.321	-0.426
x50	0.215	-0.457	0.079	-0.158	-0.363	0.105	0.067	-0.085	0.140	-0.496
x51	-0.005	0.203	-0.050	0.310	0.499	-0.154	0.041	0.013	-0.135	-0.012
x52	0.405	-0.604	0.028	0.190	0.030	-0.179	0.005	-0.237	0.163	-0.739
x53	-0.147	-0.417	-0.179	0.020	-0.120	-0.155	-0.011	-0.045	-0.238	-0.408
x54	-0.104	-0.178	0.310	-0.399	-0.263	-0.094	-0.354	-0.146	-0.028	-0.257
x55	0.250	-0.292	0.047	0.242	0.051	-0.249	-0.193	0.084	-0.508	-0.566

	Output_a	Output_b	Output_c	Output_d	Output_e
h1	0.563	-1.135	-1.297	-1.185	1.425
h2	-0.279	2.611	-1.179	-1.175	-2.065
h3	-0.831	-2.266	0.105	0.699	0.731
h4	-2.367	-0.299	-0.967	-1.529	1.783
h5	1.166	0.08	-1.587	-0.566	-0.274
h6	0.541	-0.805	1.893	-0.531	-1.952
h7	-2.274	2.322	-0.084	-1.102	-2.221
h8	-2.5	0.114	0.943	0.294	-1.56
h9	0.442	-0.821	1.578	-1.288	-1.241
h10	2.765	-0.474	-2.301	0.15	-1.564

20+ABC1

	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10
x1	-0.177	0.753	-0.32	-0.78	-0.675	0.161	0.414	-0.186	0.628	0.167
x2	-0.062	0.107	0.498	0.312	0.558	0.463	-0.389	-0.061	-0.338	-0.094
x3	0.73	-0.665	0.091	0.693	0.495	-0.5	-0.198	0.654	-0.494	-0.232
x4	-0.435	0.762	0.234	-1.42	-0.058	-0.526	0.312	0.68	0.607	0.148
x5	0.023	-0.324	0.087	0.596	0.146	-0.092	-0.463	-0.269	-0.126	-0.11
x6	0.26	-0.268	0.054	0.241	-0.108	0.283	0.215	-0.626	-0.144	0.158
x7	0.259	0.276	-0.193	-0.196	-0.231	-0.27	-0.156	-0.21	0.254	-0.302
x8	0.904	-0.226	1.238	0.281	0.583	-0.584	-0.029	-0.224	-0.392	0.404
x9	-0.686	-0.156	-0.601	-0.196	-0.457	0.566	-0.099	-0.262	0.705	-0.289
x10	-0.33	-0.249	0.218	-0.475	0.007	0.255	0.289	-0.153	0.006	-0.254
x11	-0.649	-0.233	-0.946	-0.242	-0.592	0.364	-0.097	-0.233	0.97	-0.016
x12	0.726	-0.283	-0.03	0.369	0.65	-0.283	0.537	0.248	-0.273	0.219
x13	0.349	-0.428	-0.044	0.14	-0.265	0.059	0.313	-0.242	0.457	-0.281
x14	0.673	-0.418	-0.216	0.441	0.315	-0.696	-0.045	0.121	0.135	0.497

x15	0.263	-0.037	-0.601	0.437	-1.068	0.056	-0.112	0.078	-0.303	-0.368
x16	0.539	-0.457	0.534	-0.147	1.136	-0.887	0.238	-0.13	-0.591	0.355
x17	0.111	-0.956	-0.111	1.407	0.15	1.155	-0.462	-0.194	0.389	-0.588
x18	-0.885	1.463	-0.392	-1.427	0.371	-0.766	0.838	0.604	-0.366	0.201
x19	0.429	-0.457	-0.823	1.689	-0.386	0.111	-0.083	0.348	0.142	-0.153
x20	-0.004	0.193	-0.604	-0.642	0.253	0.569	-0.495	0.364	-0.625	-0.462
x21	-0.405	-0.009	0.982	-0.028	-1.308	0.525	-0.098	-0.826	0.766	-0.489
x22	-0.237	0.268	0.336	-0.356	-0.133	0.102	-0.139	0.065	-0.462	0.225
x23	-0.482	-0.067	0.285	-0.176	-0.159	0.9	-0.146	-0.087	-0.137	-0.368
x24	0.106	-0.038	-0.352	-0.078	0.2	-0.692	0.353	0.163	0.419	0.461
x25	-0.056	0.642	-0.821	-1.164	-0.571	0.009	0.22	-0.032	0.178	-0.313
x26	-0.276	0.264	0.361	0.051	0.547	0.456	-0.229	-0.388	-0.104	-0.128
x27	0.797	-0.49	0.361	0.954	0.229	-0.473	0.167	0.373	-0.038	-0.069
x28	-0.78	-0.264	0.138	0.017	0.381	0.344	-0.152	-0.395	0.339	0.286
x29	-0.093	0.153	-0.241	-0.329	-0.287	-0.558	-0.179	-0.297	0.314	0.157
x30	0.089	-0.27	0.393	0.454	-0.077	-0.245	-0.089	-0.084	0.268	-0.373
x31	0.002	-0.601	0.511	-0.091	0.222	0.294	0.083	0.034	0.118	-0.29
x32	0.214	-0.068	-0.506	-0.794	0.317	0.862	-0.033	0.195	-1.033	-0.33
x33	0.468	0.474	-0.075	-0.252	-0.761	-0.481	-0.154	-0.102	-0.333	0.065
x34	-0.048	-0.128	0.248	-0.286	-0.935	-0.358	0.36	0.059	-0.34	-0.125
x35	-0.022	0.323	-0.491	0.02	-0.173	-0.436	-0.207	-0.128	0.213	0.139
x36	-0.217	-0.088	0.128	0.107	-0.006	-0.312	0.28	0.278	0.029	0.336
x37	-0.106	-0.251	0.383	-0.049	-0.158	0.357	0.272	0.381	0.134	0.128
x38	-0.253	0.115	0.78	0.037	-0.14	0.546	0.209	0.085	0.111	0.022
x39	0.089	0.316	-0.082	-0.236	-0.138	-0.312	0.209	0.035	-0.581	-0.224
x40	0.143	0.667	-0.03	-0.349	-0.147	0.193	0.065	-0.078	0.006	-0.211
x41	0.079	0.406	-0.122	-0.717	0.017	-0.093	0.077	-0.012	0.162	0.026
x42	0.21	0.585	-0.095	-1.05	-0.005	0.177	0.031	-0.266	0.073	0.182
x43	-0.088	0.01	-0.52	-0.228	-0.148	-0.185	0.219	0.06	0.103	-0.04
x44	0.02	-0.145	-0.113	0.014	0.143	-0.206	0.059	0.125	0.13	0.232
x45	0.146	0.152	-0.187	-0.252	-0.053	0.177	0.264	0.177	0.343	-0.084
x46	0.035	0.238	0.079	0.004	0.084	-0.065	-0.132	0.124	-0.029	0.003
x47	0.103	0.017	-0.17	0.131	0.018	-0.144	-0.005	0.059	-0.012	-0.095
x48	-0.174	-0.169	0.026	-0.037	0.201	0.138	-0.006	-0.123	-0.237	0.209
x49	-0.304	0.513	0.283	0.019	0.332	0.503	0.201	0.159	-0.131	-0.171
x50	-0.294	-0.059	-0.342	0.538	0.239	-0.034	0.228	-0.241	0.01	0.132
x51	0.403	-0.344	-0.248	-0.101	-0.034	-0.169	0.154	-0.153	0.238	0.163
x52	-0.149	0.77	-0.467	-0.445	0.195	-0.1	-0.277	-0.085	0.066	0.055
x53	-0.168	0.819	0.038	-0.019	-0.115	-0.086	0.055	0.011	-0.092	-0.21
x54	0.077	-0.172	0.022	-0.215	-0.385	0.06	-0.309	0.231	-0.209	-0.28

x55	-0.328	0.319	-0.415	-0.495	-0.364	-0.055	0.434	0.194	0.645	0.21

	Output_a	Output_b	Output_c	Output_d	Output_e
h1	-1.024	1.61	-1.089	-0.633	-1.964
h2	-1.594	-1.248	1.172	0.907	0.895
h3	-1.475	-0.054	3.028	-1.684	-2.049
h4	1.547	1.833	-3.145	-1.694	-1.746
h5	-2.678	2.105	0.109	-0.571	-1.197
h6	1.886	-1.796	0.487	0.599	-2.748
h7	-1.783	0.359	-0.403	-0.509	1.394
h8	-1.856	-0.878	-0.592	0.574	0.928
h9	0.871	-1.873	-1.089	-0.947	1.226
h10	-0.873	0.414	-0.997	-1.344	-0.231