T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
EKONOMETRİ PROGRAMI
YÜKSEK LİSANS TEZİ

METİN MADENCİLİĞİNDE KATEGORİK DEĞİŞKENLER İÇİN BENZETİM KATSAYILARININ KULLANILMASI ÜZERİNE BİR ÇALIŞMA

Emine Eda ÇAM

Danışman
Prof. DR. Levent ŞENYAY

İZMİR 2019

YÜKSEK LİSANS
TEZ ONAY SAYFASI

Üniversite	: Dokuz Eylül Üniversitesi
Enstitü	: Sosyal Bilimler Enstitüsü
Adı ve Soyadı	: EMİNE EDA ÇAM
Öğrenci No	: 2016800751
Tez Başlığı	: Metin Madenciliğinde Kategorik Değişkenler İçin Benzetim Katsayılarının Kullanımı Üzerine Bir Çalışma
Savunma Tarihi	: 17/07/2019
Danışmanı	: Prof.Dr.Levent ŞENYAY

JÜRİ ÜYELERİ

Ünvanı, Adı, Soyadı	**Üniversitesi**	**İmza**
Prof.Dr.Levent ŞENYAY	-Dokuz Eylül Üniversitesi	.....
Prof.Dr.Cenk ÖZLER	-Dokuz Eylül Üniversitesi	.....
Prof.Dr.Şaban EREN	- Yaşar Üniversitesi	.....

EMİNE EDA ÇAM tarafından hazırlanmış ve sunulmuş olan bu tez savunmada başarılı bulunarak oy birliği(X) / oy çokluğu( ) ile kabul edilmiştir.

Prof. Dr. Metin ARIKAN
Müdür

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum “Metin Madenciliğinde Kategorik Değişkenler İçin Benzetim Katsayılarının Kullanılması Üzerine Bir Çalışma” adlı çalışmanın, tarafımdan, bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

Tarih

..../..../.....

ÖZET

Yüksek Lisans Tezi

Metin Madenciliğinde Kategorik Değişkenler İçin Benzetim Katsayılarının Kullanılması Üzerine Bir Çalışma

Emine Eda Çam

Dokuz Eylül Üniversitesi

Sosyal Bilimler Enstitüsü

Ekonometri Anabilim Dalı

Ekonometri Programı

Metinlerin sayısal ortamlara aktarılması teknolojik gelişmeler yardımı ile kolaylaşmıştır. Metinlere ulaşılabilmesi ve metinlerden anlamlı bilgilerin elde edilebilmesi için metin madenciliği tekniklerinin uygulanması gerekmektedir. Metin madenciliği, yapısal formda bulunmayan verilerden doğal dil işleme disiplini ile birlikte metin üzerinden istatistiksel sonuçları elde etmeyi amaçlamaktadır.

Bu çalışma kapsamında, 4 farklı yazarın 6 farklı eseri üzerinde metin madenciliği teknikleri ve benzerlik-uzaklık ölçüleri kullanılarak R programı yardımıyla yazar tanıma ve metin sınıflandırma başarısı elde edilmeye çalışılmıştır. Yazarların eserleri ile ilgili sayısal veriler çıkartılarak bu sayısal veriler ile arasında Öklid uzaklığı hesaplanmış olup K-NN algoritması yardımıyla metin sınıflandırma yapılmıştır. Yazarı belli olmayan bir eserin elde edilen verilerden hareket ile hangi metin kümesine ait olduğu tespit edilmeye çalışılmıştır.

Anahtar Kelimeler: Metin Madenciliği, Yazar Tanıma, Benzerlik Ölçüleri.

ABSTRACT

Master's Thesis

A Study On The Use Of Simulation Coefficients For Categorical Variables In Text Mining

Emine Eda ÇAM

Dokuz Eylül University

Graduate School of Social Sciences

Department of Econometrics Program

The transfer of texts into digital environments is facilitated by technological developments. Text mining techniques should be applied in order to reach the texts and to obtain meaningful information from the texts. Text mining aims to obtain statistical results over the text along with the natural language processing discipline from data that is not found in the structure form.

Within the scope of this study, by using the text mining techniques and similarity-distance measurements on 6 different works of 4 different authors, the success of author recognition and text classification with the help of R program has been tried to be achieved. The numerical data related to the authors' works were extracted and the Euclidean distance was calculated between these numerical data and the text classification was done by using the K-NN algorithm. It has been tried to determine which text set belongs to a given work, based on the data obtained.

Keywords: Text Mining, Author Recognition, Similarity Measures.

METİN MADENCİLİĞİNDE KATEGORİK DEĞİŞKENLER İÇİN BENZETİM KATSAYILARININ KULLANILMASI ÜZERİNE BİR ÇALIŞMA

İÇİNDEKİLER

ÜÇÜNCÜ BÖLÜM
ANALİZ VE YORUM

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

EKLER LİSTESİ

GİRİŞ

Teknolojinin hızlı gelişimi ile bilgi sayısı artmakta ve büyük veriler depolanabilir duruma gelmektedir. Doğada işlenmemiş olarak bulunan verilerden anlamlı bilgilerin çıkarılması için veri madenciliği tekniklerinden yararlanılmaktadır.

Veri madenciliği, veri ambarlarında yer alan çok fazla sayıdaki veri içerisinden anlamlı bilgilerin ve bu bilgiler arasındaki ilişkilerin çıkarılmasında ve elde edilen sonuçlardan yola çıkılarak gelecekle ilgili tahminde bulunulmasını sağlayan bilgi keşfi sürecidir. Veri madenciliği yapısal veriler yani sayısal veriler ile ilgilenmektedir.

Veri madenciliği, müşterilerin satın alma tercihlerinin belirlenmesinde, dolandırıcılık tespitlerinde, müşterilerin ödeme performanslarının belirlenmesinde, pazar sepeti analizinde, tıp, biyoloji vb. alanlarda sıkça kullanılmaktadır.

Metin madenciliği ise veri madenciliğinin tersine yapısal olmayan veriler ile ilgilenmektedir. Yapısal olmayan veri türlerine resim dosyalar, pdf, word ve text gibi metin dosyalar, e-postalar örnek verilebilmektedir.

Metin madenciliği, yapısal olmayan veri setleri sayısallaştırıp yapısal hale dönüştürüldükten sonra veri madenciliği teknikleri kullanılarak elde edilen verilerden anlamlı bilgilerin çıkarılması işlemidir.

Metin madenciliği, metinler arasındaki benzerliklerin bulunması, metinlerin sınıflandırılması, metinlerin özetlenmesini sağlamaktadır.

Yapılan çalışma kapsamında metin madenciliği tekniklerinden yararlanılarak ele alınan 24 metin özetlenerek metinler hakkında bilgi çıkarımı, yazarlık özelliklerinin çıkarılması ve benzerlik ölçülerinin yardımıyla metinlerin sınıflandırılması ve sınıflandırma başarıları tespit edilmiştir. Yazarlık özellikleri ve benzerlik ölçüleri yardımıyla yazarı belli olmayan bir eserin yazarının tahminlenmesi yapılmıştır.

Tez çalışması 3 bölümden oluşmaktadır. İlk bölümde veri madenciliği kavramı, tarihsel gelişimi, veri madenciliği işlemlerinde karşılaşılan problemler, veri madenciliğinde kullanılan algoritmalar ve sınıflandırma başarı ölçümlerine yer verilmiştir.

Tezin ikinci bölümünde metin madenciliği anlatılmış olup, Metin madenciliği uygulama alanları ve uygulama adımları, Birliktelik Kuralı, Benzerlik ve Uzaklık Ölçüleri, Kümeleme Analizi, Temel Bileşenler Analizi ve Duygu Analizi ile Yazar Tanıma konuları anlatılmıştır.

Üçüncü ve son bölüm olan Analiz ve Yorum bölümünde ise, tez kapsamında yapılan deneysel çalışmaya yer verilmiştir. 4 farklı yazarın 6 farklı eseri ele alınarak 24 metinden oluşan bir veri kümesi oluşturulmuştur. Metin madenciliği ön işleme aşamaları R programı aracılığı ile metinlere uygulanmıştır. Her yazara ait yazarlık özellikleri çıkarılmıştır. R programı aracılığı ile metinler ile ilgili çıkarılan özellikler Öklid benzerlik ölçüsü yardımıyla sınıflandırılmıştır. Duygu analizi çalışması yapılmıştır.

BİRİNCİ BÖLÜM
VERİ MADENCİLİĞİ

1.1. TARİHSEL GELİŞİMİ

1950'li yıllarda ilk bilgisayarların sayma ve temek aritmetik işlemler için kullanılmaya başlanmasını takip eden 1960'larda bilgi miktarındaki artış ve bilgiyi depolama ihtiyacı ortaya çıkmaya başlamıştır. Bununla birlikte veri tabanı ve bilgilerin depolanması kavramı bilişim dünyasında yerini almıştır. Bu süreci 1970' lerde E. F. Codd' un yayınladığı ilişkisel veri tabanları makalesi sayesinde bilim dünyası veri tabanları, ilişkisel veri tabanları gibi konular ile tanışması sağlanmıştır.

1980'lerde SQL standart dil olarak kabul edilirken, 1990'larda Visual Basic, ODBC, Excel ve Access dilleri ile de veriler işlenmiştir. Böylece 1990'lı yıllarda istatistik ve veri madenciliği de ortak bir alanda buluşmuştur. Veri madenciliği sürecinin kökleri 1960'lara dayanırken ilk yazılım 1992 yılında geliştirilmiştir. Veri madenciliği, 1990'lı yılların başından bu yana sürekli gelişmektedir. Veri madenciliğinin tarihsel gelişimi özet tablo halinde aşağıda gösterilmiştir.

Tablo 1: Veri Madenciliğinin Tarihsel Gelişimi

Tarih	Basamaklar	Sorular	Kullanılabilir Teknolojiler	İlgili Yazılımlar
1960lar	Veri toplama, Veritabanı Yönetim Sistemleri	Benim son 5 yıldaki toplam kârım nedir?	Bilgisayar, Disk, Düz dosyalar	Fortran
1980ler	Veriye ulaşım, Veri sorgulama	Geçen Mart İstanbul'daki birim satış miktarı nedir?	Daha hızlı ve ucuz bilgisayarlar, daha fazla depolama alanı, ilişkisel veritabanları	Oracle,IBM DB, SQL
1990lar	Veri ambarları, Karar destek sistemleri	Geçen Mart İstanbul'daki birim satış miktarı nedir? Ankara ile karşılaştırmalı olarak görmek istiyorum.	Daha hızlı ve ucuz bilgisayarlar, Daha fazla depolama alanı, İlişkisel veritabanları, OLAP, Çok boyutlu veritabanları, Veri ambarları	SQL Standart, Veri Ambarları, OLAP, Darwin, IBM Intelligent Miner, SPSS Crisp DM, SAS Miner, Angoss Knowledge Seker
1990ların sonu 2000ler	Veri madenciliği Web madenciliği	Ankara'da gelecek ayki birim satışlarım ne durumda olacak ? Neden?	Daha hızlı ve ucuz bilgisayarlar, Daha fazla depolama alanı, İlişkisel veritabanları, Gelişmiş bilgisayar algoritmaları	Oracle Data Miner, IBM DB2 UDB Mining, SPSS Clementine, SAS Enterprise Miner

Kaynak: Kocamış ve Durmaz, 2008: 4.

1.1.2. Veri Madenciliği

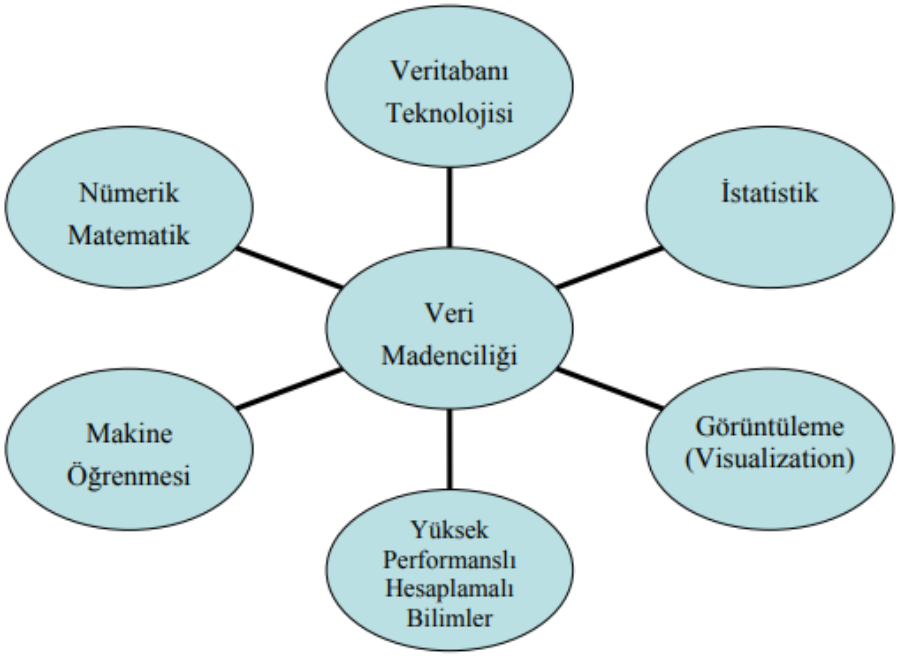
Günümüzde bilgiye hızlı erişim ve artan bilgi sayısı sebebiyle büyük veri setleri ortaya çıkmaktadır. Bu büyük veri setlerinin analizi ve yorumlanmasının zorluğu ortaya veri madenciliği sürecini ortaya çıkartmaktadır. "Veri madenciliği; veri özetleme, birliktelik kuralları, karar ağaçları, yapay sinir ağları, sınıflama algoritmaları gibi teknikleri kullanarak, büyük veri setleri arasındaki önceden bilinmeyen, geçerli ve uygulanabilir bilginin veri yığınlarından dinamik bir süreç ile elde edilmesi olarak tanımlanabilir. Diğer bir deyişle,

Veri madenciliği, elde bulunan verilerden yola çıkarak gelecekte ortaya çıkabilecek gizli örüntüleri elde etme sürecidir.

Veri madenciliği büyük miktarda veri inceleme amacı üzerine kurulmuş olduğu için bilgisayar veri tabanları ile yakından ilişkilidir.

Veri madenciliği işlemleri, sınıflama, kategori etme, tahmin etme ve görüntüleme olmak üzere 4 ana başlıkta toplanabilir. Veri madenciliği istatistik, makine bilgisi, yapay zeka, veritabanları ve yüksek performanslı işlem gibi kavramlar ile büyük ilişkiler içermektedir.

Şekil 1: Veri Madenciliği İle İlişkili Kavramlar



Kaynak: Kocamış ve Durmaz, 2008: 3.

Veri madenciliği çok sayıda veri ve çok sayıda değişken ile ilgilenmektedir.

1.1.2. Veri Ambarları

Veri ambarları farklı operasyonel sistemler, çağrı merkezleri gibi kaynaklardan veriyi alıp, temizleyip, değiştirdikten sonra mevcut veriyi daha anlaşılabilir ve daha kolay erişilebilir bir yapıda toplayan ve geçmiş veriler için bir depo teşkil etmek amacıyla oluşturulan ilişkisel veri tabanlarıdır. Sorgulama

ve analiz amacıyla kullanılmasından dolayı veri tabanlarından ayrılan yönleri bulunmaktadır. Veri ambarları, büyük veri setlerine sahip veri tabanlarından kaynaklanan iş yükü ile, analiz yükünü birbirinden ayırması ile verinin daha kolay işlenebilmesi, analiz edilebilmesi ve bu sayede veri tabanlarının daha kolay organize edilmesine olanak sağlamaktadırlar. Veri tabanlarının daha özel bir hali gibi düşünülebilir. "Veri ambarları, veri tabanlarına göre daha hızlı çalışan, özelleştirilmiş ve daha az veri saklayan yapıdadırlar. Eldeki veriyi kolay, hızlı ve doğru bir şekilde analiz etmek için gerekli olan işlemleri yerine getirmektedir." (Şeker, 2015: 6) Veri ambarları, veri tabanlarındaki değişimlere bağlı olarak güncellenebildiği için veri tabanlarından beslenir şeklinde de yorumlanabilmektedir. Veri tabanlarındaki büyük veri setlerinden daha özet verileri alır bunları amaca yönelik ve daha verimli kullanılabilmek adına saklı tutar.

Veri ambarı, 1991 yılında ilk kez William H. Inmon tarafından yönetimin kararlarını desteklemek amacı ile çeşitli kaynaklardan elde ettikleri bilgileri zaman değişkeni kullanarak veri toplama amacıyla ortaya atılmıştır. Kısaca birçok veritabanından alınarak birleştirilen verilerin toplandığı depolardır.

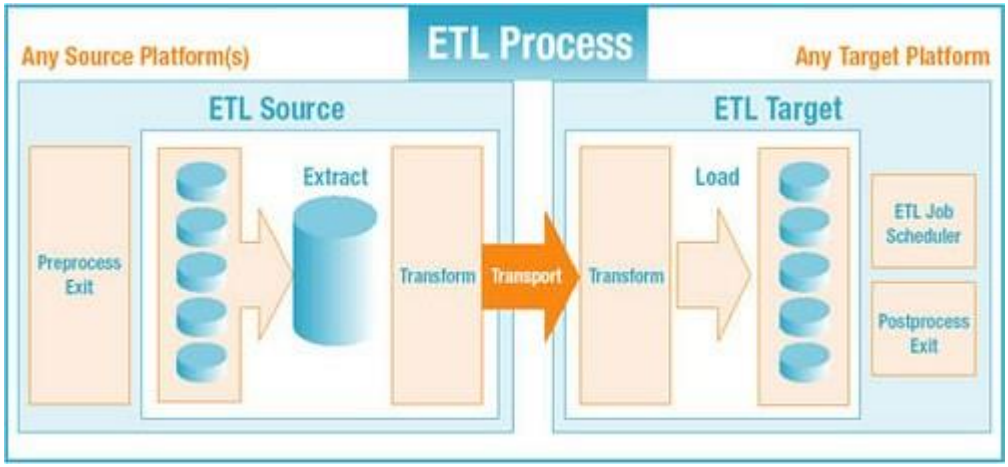
Veri ambarlarında veriler okumaya yönelik olarak oluşturuldukları için veriler, analiz yapmayı kolaylaştıran bir formatta tutulmaktadır. Burada analiz; sorgular, raporlar, karar destek sistemleri veya istatistiki hesapları kapsamaktadır. Veriler, veri ambarlarına aktarılmadan önce bir takım ön işlemlerden geçmektedir. Ön işleme adımları Bölüm 2 de çalışılmıştır. Verilerin, veri ambarlarına alınmadan önce aşamalardan geçtiği bu sürece ETL süreci denilmektedir. (Extract, Transform, Load)

Extract: Veriyi kaynak sistemden alma,

Transform: Verilerin bizim belirlediğimiz yapıya uygun olabilmesi için belli bir dönüşümlerden geçmesi gerekmektedir. Kısaca verinin temizlenmesi ve kalitesinin arttırılması,

Load: Verilerin hedef sisteme yüklenmesi anlamına gelmektedir.

Şekil 2: ETL Süreci



1.2. VERİ MADENCİLİĞİ UYGULAMA ALANLARI

Günümüzde bilişim sektöründeki gelişmelerle birlikte çok fazla veri ile karşılaşmaktayız. Gerek bilişim sektörü gerekse bankacılık, sağlık sektöründe büyük veri ile çalışılmaktadır. Bu yüzden de veri madenciliği büyük veri kümelerinin bulunduğu her ortamda kendisine çalışma alanı bulabilir.

1.2.1. Pazarlama

- Müşterilerin satın alma alışkanlıklarının belirlenmesinde
- Sepet analizi
- Müşterilerin demografik özelliklerinin çıkarılmasında
- Bütün müşterilerin e-mail, işlem, çağrı merkezi ve anket gibi erişim noktalarından elde edilen bilgiler sayesinde müşteri değerleme ve müşteri ilişkileri yönetimi
- Satış analizi yapılarak mevcut satışların arttırılmasında

1.2.2. Sağlık

- Geçmiş hasta verilerinden yola çıkarak risk grubunun belirlenmesinde
- Test verilerinden elde edilen sonuçlar ile bazı hastalıkların ön tanısı elde edilebilir.
- Ürün geliştirme

"Örneğin, Vysis, ilaç sanayisi için protein analizini yapay sinir ağları algoritmasını kullanarak yapmaktadır. Rochester Kanser Merkezi bölümü araştırmalarında Knowledge SEEKER adlı karar ağacı tekniğini kullanır. California Hastanesi veriden bilgi üretmek için "Information Discovery" adlı ürünü kullanmaktadır. Hastanede çalışan bir doktor, bu program sayesinde hastalarını hiçbir fiziksel teste tabii tutmadan teşhis koyabildiğini açıklamıştır." (Baykasoğlu, 2005: 8)

1.2.3. Bankacılık

- Müşteri profili belirlenerek müşteri kredibilitesi ve risk durumu belirlenir.
- Müşteri geçmiş dönem davranışları ele alınarak dolandırıcılık tespiti yapılabilir. Ercan Özbay'ın 2007 yılında yapmış olduğu Finans Sektöründe Veri Madenciliği ile Dolandırıcılık Tespiti adlı çalışması veri madenciliğinin finans ve dolandırıcılık tespitinde nasıl kullanıldığını anlatan örnek bir çalışmadır.
- Kar analizi
- Portföy yönetimi

1.2.4. Doküman Verileri

"Doküman veri madenciliğinde (text mining) ana amaç dokümanlar arasında ayrıca elle bir sınıflama gerekmeden benzerlik hesaplayabilmektir. Bu genelde otomatik olarak çıkarılan anahtar kelimelerin tekrar sayısı sayesinde yapılmaktadır. Polis ve emniyet, güvenlik kayıtlarında mevcut rapora benzer kaç adet ve hangi raporlar var. Ürün tasarım dokümanları ve internet dokümanları arasında mevcut tasarım için kullanılabilecek ne tür dosyalar var gibi sorulara yanıt bulunabilir. " (Baykasoğlu, 2005: 9)

1.2.5. Endüstri

- Kalite kontrol analizlerinde
- Lojistik
- Üretim planlama

- Üretim sistemlerinin benzetiminden elde edilen verilerden sistem performansını etkileyen faktörlerin ve kuralların çıkarılması alanları sayılabilir.

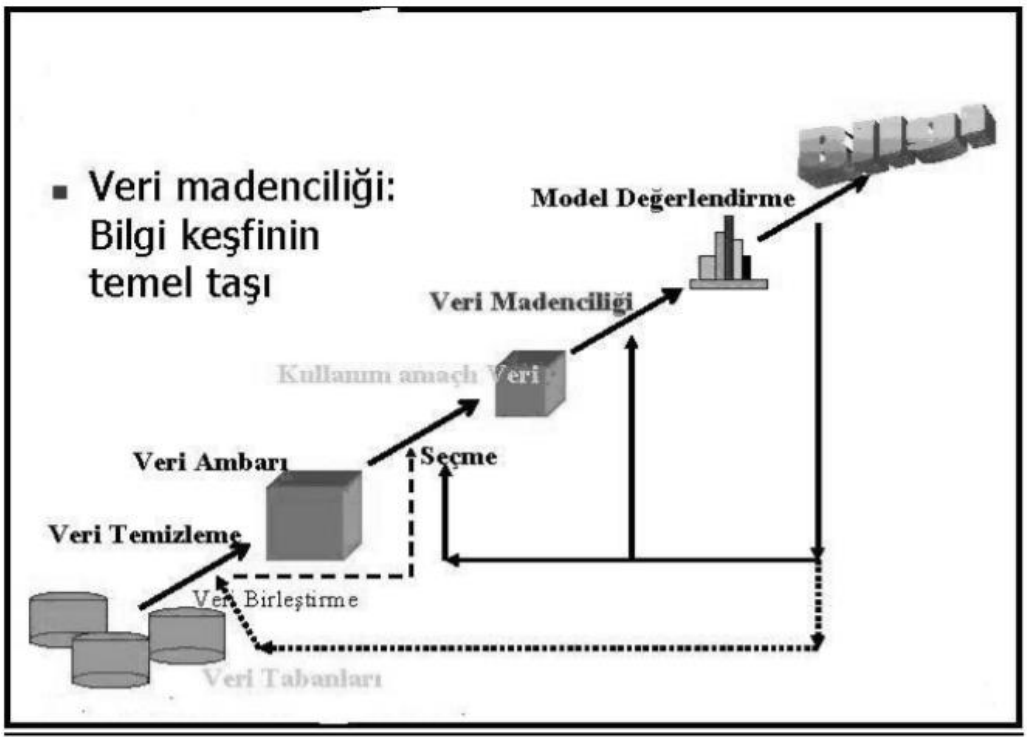
1.3. VERİ MADENCİLİĞİNDE BİLGİ KEŞFİ SÜRECİ

Veri madenciliği, bilgi keşfinin bir alt sürecidir. Veriler işlendiği takdirde bilgiye dönüşür. Veri madenciliğinin amacı olan ham verilerden yola çıkarak onlardan anlamlı ve gizli örüntüleri etmek bilgi keşfi ile gerçekleşmektedir. Bilgi keşfinin temel aşamaları aşağıdaki gibi gösterilmektedir.

- Problemin Tanımlanması,
- Verilerin Hazırlanması,
- Modelin Kurulması ve Değerlendirilmesi,
- Modelin Kullanılması,
- Modelin izlenmesi

Bilgi keşfi sürecindeki en önemli nokta veri ambarlarındaki verilerin düzenli olmasıdır.

Şekil 3: Bilgi Keşfi Aşamaları



Kaynak: Baykasl, 2006: 95.

1.4. VERİ MADENCİLİĞİNDE KARŞILAŞILAN PROBLEMLER

Veri madenciliği, büyük ve gerçek veri setleri ile çalıştığı için bu verilerin doğru ve anlaşılır bir şekilde yapılandırılması gerekmektedir. Veri setlerinde meydana gelen veri boyutunun belirlenmesi, gürültülü veri, eksik ve artık veri, boş değerler gibi sorunlar veri madenciliği sürecini olumsuz etkilemekte ve doğru, güvenilir sonuçlara ulaşılmasına engel teşkil etmektedir. Veri analizin de gerçekçi sonuçlara ulaşabilmek için öncelikle veri setimizde mevcut bulunan bu tip sorunları minimuma indirmek hatta mümkünse ortadan kaldırmak gerekmektedir.

1.4.1. Veri Boyutu

Veri tabanlarının boyutları bazı uygulamalarda az iken bazı uygulamalarda çok fazla olabilmektedir. "Veri setlerinde bulunan veri sayısının çok fazla veya çok az olması ya da nitelik sayısının çok ya da az olması algoritmalar için problem yaratmaktadır. Buna ek olarak veri setlerindeki veri sayısının zaman içinde değişkenlik göstermesi yapılan çalışma sonuçları üzerinde farklılaşmış çıktılara neden olmaktadır." (Aydilek, 2013: 18)

"Veri tabanlarında bulunan veriler yatay ve dikey olarak artmaktadır. Yatay boyutta veriler nesnelerin özellik sayılarına göre artmaktayken, dikey boyutta ise veri sayısına göre artmaktadır. Bu sorunlara çözüm olarak,

a.) Eğitim kümesinin yatay ve dikey boyutta indirgenmesi,

- ✓ Yatay indirgeme: Nitelik değerleri genelleştirilerek var olan nitelik değeri bu genelleştirilen değere göre günlenir. Mükerrer kayıtlar çıkarılır yani artık bir niteliğin değer sayısı mesela gün 24 saat ile ifade edilirken 3 parçaya bölünüp eski değerler bunlara göre günlenebilir.
- ✓ Dikey indirgeme: Artık niteliklerin indirgenmesi işlemidir.

b.) "Veri Madenciliği yöntemleri sezgisel bir yaklaşımla arama uzayını taramalıdır." (Özbay, 2007: 44)

1.4.2. Gürültülü Veri

"Gürültülü veri, veri değerleri içerisinde bazı değerlerin gerçeği yansıtmaması durumunda karşılaşılan yanlış veri değerleridir ve sistemsel olarak bir niteliğe ait tüm değerlerde bulunabilmektedir. Gürültülü veriler, veri toplanması veya veri girişi sırasında meydana gelmektedir. Bu tür verilerin veri kümesinin çalışma konusuna göre tespit edilmesinin zor olduğu durumlar olabilmektedir. Eğer gürültü, örneğin insan yaş nitelik değeri 1000 olarak kaydedilmiş gibi durumlarda tespit edildiği zaman tutarsız veri olarak da adlandırılmaktadır. Her gürültülü veri tutarsız değerler göstermeyebilmekte fakat her tutarsız veri gürültü olarak adlandırılabilmektedir. Sonuç olarak bu türdeki veri değerleri veri kümelerinde istenmeyen yanlış değerli veriler olarak görülmektedir. " (Aydilek, 2013: 18)

"Gürültülü verinin sebep olduğu problemler tümevarımsal karar ağaçlarında uygulanan yöntemler bağlamında kapsamlı bir şekilde araştırılmıştır. Eğer veri kümesi gürültülü ise sistem bozuk veriyi tanımalı ve ihmal etmelidir." (Tiryaki, 2006: 28)

Gürültülü veriyi ortadan kaldırmak için çeşitli yöntemler vardır. Bunlar;

1. Regresyon yöntemi
2. Kümeleme yöntemi
3. Kutulama (Binnig) yöntemi
4. Bilgisayar ya da insan kontrolü

1.4.3. Boş Değerler

İlişkisel veri tabanlarında sıkça karşımıza çıkan boş değer sorunu veri madenciliği sürecini olumsuz etkilemektedir. Bir veri tabanında boş değer, birincil anahtarda yer almayan herhangi bir niteliğin değeri olabilir. Boş değer, tanımı gereği kendisi de dahil olmak üzere hiç bir değere eşit olmayan değerdir.

Veri kümelerinde var olan boş değerler için çeşitli çözümler söz konusudur:

1. Boş değerli kayıtlar tamamıyla ihmal edilebilir,

2. Boş değerli kayıtlardaki boş değerler olası bir değerle güncellenebilir.
   Bu güncelleme için çeşitli yöntemler söz konusudur:

- ❖ Boş değer yerine o nitelikteki en fazla frekansa sahip bir değer veya ortalama bir değer konulabilir,
- ❖ Boş değer yerine varsayılan bir değer konulabilir,
- ❖ Boş değerinin bulunduğu kaydın diğer özelliklerine göre, NULL değerinin kendine en yakın değerle güncellenmesi sağlanabilir vb.

1.4.4. Eksik Ve Artık Değerler

Veri madenciliği çalışmalarında en önemli noktalardan biri eksiksiz veri setleri ile çalışmaktır. Veri setinin tam olması, sonuçların güvenilir ve doğru çıkmasında etkili rol oynar. Her ne kadar bu konuya özen gösterilse de veri madenciliği çalışmaları büyük veri setleri ile yapıldığı için bu gibi sorunlar ile karşılaşılmaktadır. Açık uçlu, çoktan seçmeli uzun anketlerde katılımcıların soruları atlaması, cevaplandırmaması hatta kişisel sorularda (aldatma, boşanma, cinsellik vb.) katılımcıların sorulara cevap vermemesi gibi durumlar eksik verilere sebep olmaktadır. "Eksik değerler, nitelik değerinin veri kümesine kayıt edilememesi gibi durumunda oluşabileceği gibi aykırı, gürültülü, tutarsız değerlerin tespit edilip veri kümesinden silinmesi sonucunda da oluşabilmektedir."(Aydilek, 2013: 19)

Bir veri setinde eksik değerler olduğu gibi konu ile alakası olmayan, veri kümesinin yapısını bozan veriler de bulunabilmektedir. Bu verilere `artık veri` adı verilmektedir. Eksik ve artık verilerin yok edilmesi ya da veri seti içerisindeki olumsuz etkisinin azaltılması için bir takım yöntemler geliştirilmiştir. Bunlar aşağıda sıralanmıştır,

1. Liste boyunca silme
2. Eşlerin silinmesi
3. Ortalama yöntemi
4. Doğrusal interpolasyon yöntemi
5. Son gözlemin taşınması
6. Maksimum beklenti yöntemi
7. Jackknife yöntemi

1.4.4.1. Liste Boyunca Silme Yöntemi

Liste boyunca silme yöntemi, bir veri kümesinde eksik verinin bulunduğu satırı tamamen silmektir. Veri analizi yaparken, büyük veri setleri ile çalışılıyorsa ve eksik veri sayısı çok fazla ise kullanılmaması gereken bir yöntemdir. Çünkü liste boyunca silme yöntemi, veri seti için gerekli olan yararlı verilerin de silinmesine sebep olabilmektedir. Eğer veri setimizde eksik gözlem sayısı az ise bu yöntemi kullanabiliriz.

Bir örnek ile açıklayacak olursak;

Tablo 2: Liste Boyunca Silme Yöntemi

Çiçek adı	Anavatanı	Familyası
Açelya	Kuzey Amerika ve Asya	Fundagiller
Afrika Menekşesi	Tanzanya	------
Atatürk Çiçeği	Meksika ve Orta Amerika	Sütleğengiller
Cam Güzeli	Doğu Afrika	-------
Altın Çanak	------	Zeytingiller
Atlas Çiçeği	Orta ve Güney Amerika	Kaktüsgiller

Yukarıdaki örneğe eksik verilerin liste boyunca silinmesi yöntemi uygulanırsa,

Çiçek adı	Anavatanı	Familyası
Açelya	Kuzey Amerika ve Asya	Fundagiller
Atatürk Çiçeği	Meksika ve Orta Amerika	Sütleğengiller
Atlas Çiçeği	Orta ve Güney Amerika	Kaktüsgiller

halini alır. Fakat daha öncede belirtildiği üzere veri sayısının, eksik gözlemlerin çok olduğu veri setlerinde bu yöntem olumsuz sonuçlara yol açabilmektedir.

1.4.4.2. Eşlerin Silinmesi Yöntemi

Eşlerin silinmesi yönteminde ise, amaca yönelik olarak işlem yapılmasıdır. Amaçladığımız veri kümesi kalırken diğer veri kümesinin göz ardı edilmesine dayanmaktadır. Yukarıdaki örneği temel alacak olursak,

Tablo 3: Eşlerin Silinmesi Yöntemi

Çiçek adı	Anavatanı	Familyası
Açelya	Kuzey Amerika ve Asya	Fundagiller
Afrika Menekşesi	Tanzanya	------
Atatürk Çiçeği	Meksika ve Orta Amerika	Sütleğengiller
Cam Güzeli	Doğu Afrika	-------
Altın Çanak	------	Zeytingiller
Atlas Çiçeği	Orta ve Güney Amerika	Kaktüsgiller

Eşlerin silinmesi yöntemi uygulanırsa,

Diyelim ki bulmak istediğimiz çiçek adı ve anavatanı eşleştirmesi olsun. Bu amaç doğrultusunda familyalarını göz ardı ederek sadece anavatanı veri kümesinde eksik olan veri çıkarılır.

Çiçek adı	Anavatanı	Familyası
Açelya	Kuzey Amerika ve Asya	Fundagiller
Afrika Menekşesi	Tanzanya	------
Atatürk Çiçeği	Meksika ve Orta Amerika	Sütleğengiller
Cam Güzeli	Doğu Afrika	-------
Atlas Çiçeği	Orta ve Güney Amerika	Kaktüsgiller

Tabi bu yönteminde eksik kaldığı ya da her zaman doğru sonuçlar üretemeyeceği durumlar vardır. Her işlem için farklı veri kümesi üretiliyor olması

ve hatta her işlem sonucunda farklı sayıda veri ele alınıyor olması nedeniyle işlemler sonucunda verilerin karşılaştırılma güçlüğü ortaya çıkabilmektedir.

1.4.4.3. Ortalama Yöntemi

Ortalama yönteminde eksik veri, verinin ait olduğu veri kümesinin ortalaması alınarak bulunabilmektedir. Bu yöntem gözlem sayısının çok olduğu veri setlerinde sağlıklı sonuçlar vermemektedir. Ayrıca ortalama yöntemi sadece sayısal verilerin olduğu ver setlerinde geçerli olup, nominal değerlerin bulunduğu veri setlerinde geçerli değildir. Ortalama yöntemi örnek verecek olursak,

Tablo 4: Ortalama Yöntemi

Çalışan adı	Departman	Maaş
Kemal Bozkurt	İnsan Kaynakları	3000
Sibel Eker	Pazarlama	2500
Nihat Kuru	Bilgi İşlem	------
Nuray Tarhan	Muhasebe	3000

Eksik gözlemin yer aldığı maaş sütununda amacımız Nihat Kuru adlı çalışanın maaşını bulmak olsun.

Maaş(ortalama)= (3000+2500+3000) / 3

Maaş(ortalama)=2.833

Çalışan adı	Departman	Maaş
Kemal Bozkurt	İnsan Kaynakları	3000
Sibel Eker	Pazarlama	2500
Nihat Kuru	Bilgi İşlem	2835
Nuray Tarhan	Muhasebe	3000

2833 olarak bulunan ortalama veri setinde eksik gözlem içerisinde 2835 olarak yuvarlanabilir.

1.4.4.4. Doğrusal Interpolasyon Yöntemi

Doğrusal interpolasyon yöntemi, önceden bilinen değerlerden yola çıkarak eksik değerin tamamlanmasını sağlayan bir yöntemdir. Burada esas nokta verilerin belirli bir sınıf içerisinde doğrusal bir şekilde dağılmasıdır.

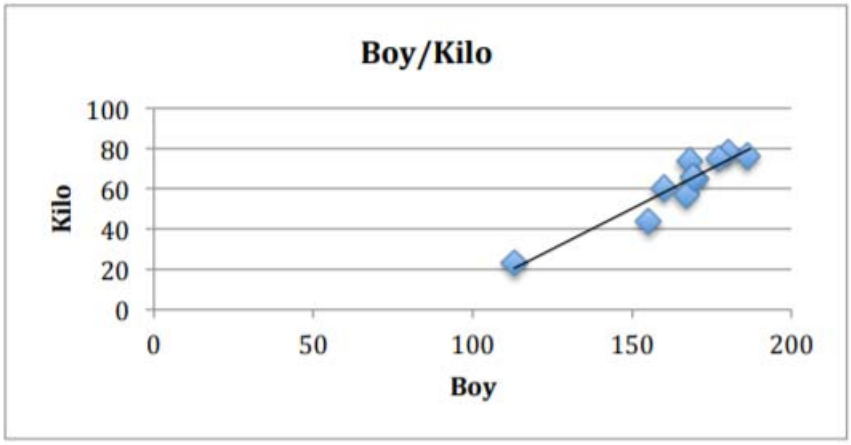
Sadi Evren Şeker ile Eda Eşmekaya' nın "Eksik verilerin tamamlanması" adlı çalışmalarında vermiş olduğu örneği ele alırsak,

Tablo 5: Boy / Kilo ve Cinsiyet Tablosu

Boy	Kilo	Cinsiyet
180	78	Erkek
177	75	Erkek
170	65	Kadın
168	74	Erkek
186	76	Kadın
187	---	Erkek
167	57	Kadın
160	60	Kadın
113	23	Erkek
169	66	Erkek
155	44	Kadın

Kaynak: Şeker ve Eşmekaya, 2017: 14)

Şekil 4: Boy/Kilo Grafiği



6.gözlemin kilo verisi kayıptır. Kayıp değeri bulabilmek için doğrusal interpolasyon yöntemi uygulanabilir. Boy değişkenini temel sınıf olarak alırsak,

$Y^* = Y_0 + (X - X_0)\frac{Y_1 - Y_0}{X_1 - X_0}$

$Y^* = 76 + (187 - 186)\frac{57 - 76}{167 - 186}$

$Y^* = 77$

Kayıp gözlem 77 olarak bulunacaktır.

1.4.4.5. Son Gözlemin Taşınması

Son gözlemin taşınması yöntemi, eksik veri setinin tamamlanabilmesinde kendisinden bir önceki gözlemden yararlanılmaktadır. Bu yöntemin olumsuz tarafı ise bazı marjinal noktaların çoğaltılması olarak görülebilir. Örneğin maaş sütununda marjinal olarak 25.000 TL maaş alan bir çalışan varsa bu kişinin kopyalanması, veri madenciliği sonuçlarını olumsuz etkileyebilmektedir.

Tablo 6: Son Gözlemin Taşınması

Ad-Soyad	Departman	Maaş
Kemal Bozkurt	İnsan Kaynakları	3000
Sibel Eker	Pazarlama	2500
Nihat Kuru	Bilgi İşlem	2835
Nuray Tarhan	Muhasebe	3000
Ece Durgut	İş geliştirme	----

Eğer yukarıdaki tabloya son gözlemin taşınması yöntemini uygulanmak istenirse,

Ad-Soyad	Departman	Maaş
Kemal Bozkurt	İnsan Kaynakları	3000
Sibel Eker	Pazarlama	2500
Nihat Kuru	Bilgi İşlem	2835
Nuray Tarhan	Muhasebe	3000
Ece Durgut	İş geliştirme	3000

Ece Durgut adlı çalışanın eksik olan maaş verisi kendisinden bir önce olan Nuray Tarhan'ın maaş verisi ile tamamlanmaktadır.

1.4.4.6. Maksimum Beklenti Yöntemi

"Maksimum beklenti yönteminde kayıp verilerin tahmin edilmesi, algoritmanın birden fazla veriyi aynı anda tahmin ederek bu birden fazla veriye aynı değeri atayarak gerçekleşmektedir. Algoritmada bir verinin hangi veri kümesine ait olduğu belirlenirken kesin mesafe ölçütlerini kullanmak yerine tahminsel ölçütler kullanılmaktadır. Maksimum beklenti yöntemi 2 aşamadan oluşmaktadır. Bunlar maksimum aşaması ve beklenti aşamasıdır. Beklenti adımında, gözlenen verilerin parametrelerine ait kestirimler kullanılarak bilinmeyen veri ile ilgili en iyi olasılıklar tahmin edilir. Maksimum adımında ise tahmin edilen kayıp veri yerine konularak bütün veri üzerinden maksimum olabilirlik hesaplanır ve parametrelerin yeni kestirim sonuçları elde edilir. " (Sezgin ve Çelik, 2013: 613)

Maksimum beklenti yöntemini bir örnek ile açıklayacak olursak;

n = veri kümesi

k = Bilinen veri sayısı

n-k = Kayıp veri sayısı

e = Kabul edilebilir hata miktarı (Eşik kriteri)

X^p = Ortalamadan hesaplanan kayıp verilere atadığımız keyfi değer

Adım 1

Kayıp verileri aşağıdaki formülden hesapla.

$$\frac{y' = \sum_{i=1}^{k} x_i + \sum_{j=k+1}^{n} x_j^p}{n}$$

$y' - x^p < e \; ise \; DUR \; değilse; \; x^p = y' \; adım \; 1 \; e \; git.$

$y' = \bar{x} + \overline{x^p}$

n=15

k=12

n-k= 3

e = 0.1 olsun.

n= (30,41,45,12,20,14,16,35,8,23,26, 50, ?,?,?)

$$y' = \frac{320}{15} = 21.333$$

$x^p = 20\ olsun.$

$$y' = \frac{320}{15} + \frac{20 + 20 + 20}{15} = 25.333$$

25.333 – 20 = 5.333 < 0.1 (Yanlış) Adım 1 e dön.

$$y' = \frac{320}{15} + \frac{25.333 + 25.333 + 25.333}{15} = 26.40$$

26.40 – 25.33 =1.07 < 0.1 (Yanlış) Adım 1 e dön.

$$y' = \frac{320}{15} + \frac{26.40 + 26.40 + 26.40}{15} = 26.61$$

26.61 - 26.40 =0.21 < 0.1 (Yanlış) Adım 1 e dön.

$$y' = \frac{320}{15} + \frac{26.61 + 26.61 + 26.61}{15} = 26.65$$

26.65 – 26.61 = 0.04 < 0.1 (Doğru) 4 adımda algoritma tamamlanmıştır.
Eksik değerler yerine 26.65 yazılabilir.

1.4.4.7. Jackknife Yöntemi İle Kayıp Verilerin Tahmin Edilmesi

"Jackknife yönteminin temel mantığı, veri setinde her bir gözlem değeri bir kez dışarıda bırakılarak geride kalan gözlemlerden örneklem istatistikleri hesaplanmaya dayanmaktadır." (Temel ve Erdoğan, 2011: 46) Bu yüzden maksimum beklenti yönteminden farklı olarak her bir kayıp değer için farklı tahmin değerinde bulunulmaktadır. Her bir kayıp değer ayrı iterasyonlar ile bulunmaktadır. Jackknife yöntemini bir örnek ile açıklayacak olursak;

n = Veri kümesi

k = Bilinen veri sayısı

j = Kayıp veri sayısı

e = Eşik değeri

$$y' = \frac{\sum_{i=1}^{k} x_i \ + x^p}{k+1}$$

$\left|x_p - \ y'\right| \leq e$ Doğru ise Dur değilse tekrarla.

n = (6 , 5 , 8 , 10 , 1 , ? , ?)

k = 5 j = 2 $\frac{30}{5} = 6$ ise $x^p = 6$ olsun.

$$y' = \frac{\sum_{i=1}^{5} x_i \ + x^p}{6} = \frac{30+6}{6} = 6$$

$|6 - \ 6| \leq 0.1$

Doğru bu yüzden ilk kayıp değer 6 olacaktır.

2.adım da ise diğer eksik değerin tahminlemesi yapılacaktır.

$x^p = 6$ olarak sabitlenecektir.

$$y' = \frac{\sum_{i=1}^{6} x_i \ + x^p}{7} = \frac{36+6}{7} = 6$$

$|6 - \ 6| \leq 0.1$ doğru bu yüzden 2. Kayıp değer de 6 değerini alacaktır.

1.5. VERİ MADENCİLİĞİ MODELLERİ VE KULLANILAN ALGORİTMALAR

1.5.1. Sınıflandırma Yöntemi

Veri madenciliği modelleri arasında en sık kullanılan model sınıflandırma modeli ve veri madenciliğinin temelidir. Sınıflandırma modellerinde, bir takım girdi değişkenlerine bağlı olarak eldeki girdilerin hangi sınıflara dahil olacağının tahminlenmesi yapılmaktadır. Sınıflandırma modellerinde 2 adet küme bulunmaktadır. Biri eğitim kümesi diğeri ise test kümesidir. Bu yüzden sınıflandırma işlemi 2 aşamada gerçekleştirilmektedir. İlk aşamada daha önce görülmemiş ve kategorisi belli olmayan bir örneğin eğitim verisi kategorileri yardımıyla en uygun sınıfa atanırken, ikinci aşamada test kümesi yardımıyla modelin geçerliliği test edilmektedir. En çok kullanılan sınıflandırma algoritmaları K-En Yakın Komşu Algoritması, Naive Bayes Algoritması, Karar Ağaçları, Yapay Sinir Ağları'dır.

1.5.2. K-En Yakın Komşu Algoritması

K-En yakın komşu algoritması, eğitimli öğrenme algoritmasıdır. Eğitimli öğrenme algoritması olmasının sebebi, önceden hazırlanmış eğitim kümesinden yola çıkarak eldeki verilerin dışında başka bir veri geldiğinde var olan öğrenme verisi üzerinden sınıflandırma yapmasıdır. Örneğin k = 5 için yeni bir eleman sınıflandırılmak istensin bu durumda eski sınıflandırılmış elemanlardan en yakın 5 tanesi alınır. Bu elamanlar hangi sınıfa dahilse, yeni eleman da o sınıfa dahil edilir. Mesafe hesabından genellikle Öklid mesafesi kullanılmaktadır. Bir örnek ile anlatılmak istenirse;

"Elimizde bir X = (araba sayısı= "2", ev sayısı = "3", mağaza sayısı = "0") örneği, sınıflandırmak istediğimiz bilinmeyen örnek olsun. Veri özelliklerimiz araba sayısı,ev sayısı ve mağaza sayısıdır. Sınıflar ise Zengin yaşam(Z), Orta halli yaşam(O), Alt seviye yaşam (A) dır. Öklid uzaklığı ile X sınıfının her bir sınıfa olan uzaklığı belirlenir. K=5 alındığında Öklid uzaklıkları küçükten büyüğe

sıralanır ve en küçük ilk 5 sınıf belirlenir. Bu değerler arasında en çok orta halli sınıfa ait elemanlar bulunduğundan X bilinmeyen örneği orta halli sınıfa atanır." (Doğan, 2006: 29)

Tablo 7:Öklid Uzaklığının Hesaplanması

	Sınıf	Araba Sayısı	Ev Sayısı	Mağaza Sayısı	Öklid Uzaklığı
1	Zengin Yaşam(Z)	5	10	4	$((5-2)^2+(10-3)^2+(4-0)^2)^{1/2} = 12.08$
2	Zengin Yaşam(Z)	3	5	6	$((3-2)^2+(5-3)^2+(6-0)^2)^{1/2} = 6.40$
3	Zengin Yaşam(Z)	4	8	7	$((4-2)^2+(8-3)^2+(7-0)^2)^{1/2} = 8.83$
4	Orta Halli Yaşam (O)	2	2	2	$((2-2)^2+(2-3)^2+(2-0)^2)^{1/2} = 2.24$
5	Orta Halli Yaşam (O)	2	3	2	$((2-2)^2+(3-3)^2+(2-0)^2)^{1/2} = 2$
6	Orta Halli Yaşam (O)	1	2	2	$((1-2)^2+(2-3)^2+(2-0)^2)^{1/2} = 2.45$
7	Alt Seviye Yaşam(A)	1	1	0	$((1-2)^2+(1-3)^2+(0-0)^2)^{1/2} = 2.24$
8	Alt Seviye Yaşam(A)	0	1	0	$((0-2)^2+(1-3)^2+(0-0)^2)^{1/2} = 2.83$
9	Alt Seviye Yaşam(A)	0	1	1	$((0-2)^2+(1-3)^2+(1-0)^2)^{1/2} = 3$

1.5.3. Naive Bayes Algoritması

Naive Bayes sınıflandırma algoritmasının temeli Bayes teoremine dayanan, veri setinde bulunan özelliklerin hepsini birbirinde bağımsız ve eşit önemlilik derecesinde ele alan, olasılığa bağlı bir veri madenciliği yöntemidir. Diğer yöntemlere göre veri sınıflandırma başarısı daha yüksektir. "Elimizde n adet sınıf olduğunu farz edelim, $S_1,S_2,\ldots,S_n$. Herhangi bir sınıfa ait olmayan bir veri örneği X'in, hangi sınıfa ait olduğu Naive Bayes sınıflandırıcı tarafından belirlenmektedir. Veri örneği X, verilen sınıflara ait olma olasılığı en yüksek değere sahip sınıfa atanır. Sonuç olarak, Naive Bayes sınıflandırıcı bilinmeyen örnek X'i, S_i sınıfına atar." (Doğan; 2006: 22) X örneğinin S_i sınıfında olma olasılığı;

$$P(S_i / X) = \frac{P(X / Si) / P(Si)}{P(X)}$$

Formülü yardımıyla hesaplanmaktadır. "P(X) bütün sınıflar için sabit ise, X örneğinin S_i sınıfında olma olasılığına $P(X/S_i)$ $P(S_i)$ ifadesi ile ulaşabiliriz. $P(S_i)$, her bir sınıfın olasılığı olup aşağıdaki gibi hesaplanmaktadır.

$$P(Si) = \frac{O_i}{O}$$

Burada O_i, S_i sınıfına ait eğitilen örnek sayısı, ve O ise toplam eğitilen örnek sayısıdır. “ (Doğan; 2006: 22)

Sınıfı belli olmayan bir X örneğini eğitim verisinden yola çıkarak Naive Bayes sınıflandırma algoritması ile hangi sınıfa gireceğini tahmin edilmek istenirse, aşağıdaki tabloda bilgisayar ve özellikleri yer almaktadır;

Tablo 8: Bilgisayar Ve Özellikleri

Sıra No	Sınıf	İşlemci Teknolojisi	Disk Kapasitesi	RAM Kapasitesi	Ekran Hafıza Kartı
1	A	İntel i5	1 TB	4 GB	2 GB
2	B	İntel i5	256 GB	8 GB	3 GB
3	A	Amd A9	1 TB	4 GB	2 GB
4	A	İntel i7	1 TB	16 GB	4 GB
5	B	İntel i3	500 GB	4 GB	2 GB
6	B	İntel i7	1 TB	8 GB	4 GB
7	A	İntel i5	512 GB	16 GB	4 GB
8	A	İntel i7	1 TB	16 GB	8 GB

X verisinin hangi sınıfa ait olduğunu belirleyebilmek için $P(X/S_i)P(S_i)$ değerini bulabilmemiz gerekmektedir. A sınıfı 5 elemandan, B sınıfı da 3 elemandan oluşmuştur. $P(S_i)$ olasılık değerleri eğitim sınıfından hesaplanabilir:

P(A) = 5/8 =0.625

P(B) = 3/8 =0.375

$P(X/S_i)$, i =1,2 değerlerinin koşullu olasılıkları;

P(işlemci teknolojisi = “İntel i5” | A) = 2/5 = 0.40

P(Disk kapasitesi = “1 TB” | A) = 4/5 = 0.80

P(RAM Kapasitesi = “4GB” | A) = 2/5 = 0.40

P(Ekran Hafıza Kartı = “2 GB” | A) = 2/5 = 0.40

P(işlemci teknolojisi = “İntel i5” | B) = 1/3 = 0.33

P(Disk kapasitesi = “1 TB” | B) = 1/3 = 0.33

P(RAM Kapasitesi = “4GB” | B) = 1/3 = 0.33

P(Ekran Hafıza Kartı = “2 GB” | B) = 1/3 = 0.33

Olasılık değerlerini ele alarak,

P(X|A) = 0.40*0.80*0.40*0.40 =0.051

P(X|B) = 0. 33*0.33*0.33*0.33 =0.011

P(A) P(X|A) = 0.625*0.051= 0.032

P(B) P(X|B) = 0.375*0.011= 0.004

Elde edilen sonuca göre X bilgisayarı en yüksek olasılık değerine sahip olan A sınıfına aittir.

1.5.4.Yapay Sinir Ağları

Yapay sinir ağları giriş, çıkış ve giriş ile çıkış katmanları arasında bulunan ara katman olmak üzere 3 bölümden oluşan, yapar sinir hücrelerinin farklı etkiler ile birbirlerine bağlanması sonucu ortaya çıkan yapıdır. "Giriş katmanında k adet giriş nöronu, gizli katmanda h adet nöron ve çıkış katmanında q adet nöron bulunan bir tek katmanlı ileri beslemeli sinir ağı k-h-q ağı olarak bilinir." (Şengür, Tekin, 2013; 9) "Örüntü tanıma (yüz tanıma, hareket tespiti, ses tanıma, imza tanıma, arıza analizi ve tespiti, tıp, otomasyon ve kontrol alanlarında yapay sinir ağlarından yararlanılmaktadır." (Taşova, 2011: 14)

"Yapay sinir ağlarının temeli biyolojik sinir ağlarının çalışma şekillerine benzedikleri için üstünlük açısından da biyolojik sinir ağlara benzemektedir. Bu yüzden yapay sinir ağlarının avantajları aşağıdaki gibi sıralanmıştır:

- Doğrusal olmayan yapı
- Paralellik
- Öğrenme
- Hafıza
- İlişkilendirme
- Sınıflandırma
- Genelleme
- Tahmin
- Eksik verilerde çalışma vb.

Bu çok yönlü olan avantajlarının yanında dezavantajları da mevcuttur. Bunlar; yapay sinir ağlarının eğitilmesine ve test edilmesine olanak sağlayacak olan çok büyük veri setleri ile çalışması (veri setinin büyüklüğü yapılan

çalışmaya göre değişmektedir.) ve uygulanmasının zor ve karmaşık olmasıdır. “ (Köktürk, 2012: 24)

Yapay sinir ağlarına literatürden, D. Şengür ve Ahmet Tekin’in yapmış oldukları “Yapay sinir ağları ile öğrenci mezuniyet notlarının tahmin edilmesi”, Ozan Taşova’nın “Yapay sinir ağları ile yüz tanıma” ve örnek el yazılarının eğitilmesi ile yapar sinir ağları yöntemiyle el yazılarının tanınması örnek verilmektedir.

1.5.5. Karar Ağaçları

Sınıflandırma algoritmaları arasında yorumlanması ve güvenilirlik açısından en kolay yöntem olan karar ağaçları en yaygın kullanılan yöntemlerden de birisidir. “Karar ağacı, çok sayıda kayıt içeren bir veri kümesini, bazı kuralları uygulayarak daha küçük kümelere bölmek için kullanılan bir yapıdır.” (Sezer, 2008: 26) “Bu ağaç yapısında her bir değişken bir düğüm tarafından temsil edilirken, dallar ve yapraklar ise ağaç yapısının diğer elemanlarıdır. Ağaçta en son kısım yaprak, en üst kısım ise kök olarak adlandırılmaktadır. Kök ve yapraklar arasında kalan kısımlar ise dal olarak ifade edilmektedir. Diğer bir deyişle bir ağaç yapısı; verileri içeren bir kök düğümü, iç düğümler (dallar) ve uç düğümlerden (yapraklar) olmak üzere 3 bölümden oluşmaktadır.” (Köktürk, 2012: 44)

1.6. SINIFLANDIRMA MODELLERİNİN BAŞARISININ ÖLÇÜLMESİ

Sınıflandırma yaparken yukarıda açıklanan algoritmaların başarısının ölçülmesi, hangi sınıflandırma algoritmasının daha iyi ve güvenilir sonuç verdiğinin ölçülmesi 4 ölçüt ile yapılmaktadır. Bunlar; doğruluk-hata oranı, kesinlik, duyarlılık ve F-ölçütüdür.

Tablo 9: Hata Matrisi

		Öngörülen Sınıf	
		Sınıf=1	Sınıf=0
Doğru Sınıf	Sınıf=1	a	B
	Sınıf=0	c	D

a= TP (True Positive) b= FN (False Negative) c= FP (False Positive) d= TN (True Negative)

1.6.1. Doğruluk-Hata Oranı

Bu dört yöntem arasında en basit olanı doğruluk oranıdır. Doğruluk oranı doğru sınıflandırılmış veri sayısının, toplam veri sayısına oranını verirken, hata oranı ise doğruluk oranın 1' den çıkartılması ile bulunmaktadır.

Doğruluk = $\frac{TP+TN}{TP+TN+FP+FN}$

Hata Oranı= 1 – Doğruluk Oranı

1.6.2. Kesinlik

Kesinlik = $\frac{TP}{TP+FP}$

1.6.3. Duyarlılık

Duyarlılık = $\frac{TP}{TP+FN}$

1.7.6. F Ölçütü

"Duyarlılık ve kesinlik ölçütü tek başına anlamlı sonuçlar vermemektedir. Bunun için her ikisinin harmonik ortalaması olan F-ölçütü ortaya çıkmıştır." (Coşkun ve Baykal, 2011: 54)

F-ölçütü = $\frac{2*Duyarlılık*Kesinlik}{Duyarlılık+Kesinlik}$

İKİNCİ BÖLÜM
METİN MADENCİLİĞİ

Veriler doğada çok farklı yapıda bulunmaktadır. Günümüzde teknolojinin hızlı gelişimi ve kullanımı ile birlikte birçok bilgi dijital alanda kayıt altında tutulmaktadır. Bir taraftan giderek artan bu belge yığınları içinde önemli birçok bilgi göz önünden kaybolup giderken, diğer taraftan gerektiğinde ihtiyaç duyulan değerli bilgilere ulaşmak için dokümanların içeriğinin belirlenmesi ve buna uygun sorgulanabilmesi ihtiyacı da giderek artmaktadır. Doğal dil ile yazılmış metinleri çeşitli sınıflandırma ve aşamalardan geçirerek test edilmesini sağlayan metin madenciliği yöntemi bilgiye ulaşma ve analiz etme sürecini kolaylaştırmaktadır.

Metin madenciliği, yapısal olmayan verileri kullanarak anlamlı ve nitelikli bilgiler elde etmeyi amaçlayan ve 2000'li yıllardan sonra ilginin giderek arttığı önemli bir alan haline gelmiştir. Metin madenciliği kelimelerle ilgilidir ve düzensiz halde bulunan bilgilerin yapılandırılması sürecidir. Yapısal halde bulunmayan veri kümesi ile çalıştığı için çeşitli yöntemler ile verileri sınıflandırılarak, gruplandırılarak, veriler arasındaki ilişkileri ortaya çıkarılarak bir model oluşturulur. Böylelikle metinsel veriler, veri madenciliği tekniklerinin uygulanabileceği uygun formata dönüştürülmüş olur.

"Metin yazımında belirli kuralların bulunmaması nedeniyle bilgisayar dili bunları anlayamamaktadır. Her bir metnin dili ve içerdiği anlam amaca bağlı olarak çeşitlilik göstermektedir. Yapısal olmayan bilgiden anlam çıkarmak için kullanılan geleneksel yöntemler; anahtar kelimeler veya mantıksal aramalar, istatistiksel veya olasılıksal algoritmalar, sinir ağları ve kalıp keşfedici sistemler gibi dilbilimsel olmayan yöntemlerdir ." (Çelikyay, 2010: 46)

"Metin madenciliği tekniklerine, günümüzde sanal dünyanın sosyal etkileşiminin popülerliğinden dolayı sanal ticaret konularında da rastlanılmaktadır. Ürün pazarlama ve tanıtma sosyal medyalar üzerinde daha rahat geliştirilip daha rahat incelenebilmektedir. Sosyal medyalarda sayısal ya da metin verilerini toplamak, derlemek ve analiz etme süreçleri daha kısa ve düşük maliyetli olması nedeniyle pazarlama beceri analizleri bu alanlar üzerinde daha çok yapılmaktadır. Reklam ve tüketim dünyasının ilişkisini göz önüne alarak ihtiyacın karşılığı daha net ortaya çıkacaktır." (Aydın, 2017: 22)

2.1. METİN VE VERİ MADENCİLİĞİ

Veri madenciliği, büyük ve farklı veri ambarlarındaki veriler arasındaki beklenmedik veya bilinmedik ilişkileri ortaya çıkaran bir alandır. Veri madenciliği bu bilinmedik ilişkileri ortaya çıkarırken bu verilerden hem anlaşılır hem de güvenilir bilgiler elde etmeyi amaçlar.

"Geleneksel sorgu ve raporlama tekniklerinin veri kümeleri karşısında yetersiz kalması, saklı ve işlenmemiş bilgiye olan gereksinim Veritabanlarında Bilgi Keşfi ve Veri Madenciliği gibi alanlar sayesinde anlaşılabilir ve yorumlanabilir duruma gelmiştir." (Oğuz, 2009: 8)

Metin madenciliği ise belirgin bir yapısı olmayan(yapılandırılmamış veri),metin biçimindeki veriler içerisinden gizli nitelikli bilginin çıkarılması ve düzensiz haldeki verinin yapılandırılması sürecidir. Metin madenciliği bilgiye ulaşma sürecinde veri madenciliği tekniklerini kullanır.

Veri madenciliği yapılandırılmış yani belirli bir yapısı olan, kullanıma hazır veriler topluluğu ile çalışırken metin madenciliği yapılandırılmamış, kesin olmayan veriler ile çalışır.

Hem veri madenciliğinde hem de metin madenciliğinde veri ambarlarındaki veriler arasındaki beklenmedik veya bilinmedik ilişkileri ortaya çıkarırken genel yapay zeka, makine öğrenme ve istatistik algoritmaları kullanılmaktadır.

2.2. METİN MADENCİLİĞİ UYGULAMA ALANLARI

"Metin madenciliği; ulusal güvenlik ve şirket güvenlik uygulamalarında, yasal-avukat-hukuk durumlarında, şirket finansı- iş aklı için, patent analizleri için, halkla ilişkiler- karşılaştırılabilir kurumların, işletmelerin Web sayfalarını karşılaştırma gibi pek çok çeşitli alanda uygulanabilir. " (Melek, 2012: 35)

Metin madenciliği, düzensiz halde bulunan yapısal olmayan verilerin tespit ve analizinde, bu metinlerin düzenli hale getirilip bunlardan özet bilgiler çıkarılmasında da kullanılmaktadır.

2.2.1. Müşteri İlişkileri Yönetimi (CRM)

Metin madenciliği teknikleri sayesinde insanların tüketim alışkanlıklarındaki değişmelerin belirlenmesi, müşterilerin yapmış olduğu şikayet, dilek ve önerilerinin tespiti yapılabilir. "Bütün müşterilerin e- mail, işlem, çağrı merkezi ve anket gibi erişim noktalarından elde edilen metin bilgilerinden nitelikli bilgi çıkarılır. Bu nitelikli bilgi müşterinin terk etme ve çapraz satışlarını tahmin etmek üzere kullanılır." (Dolgun ve diğerleri, 2009: 48-58).

Örneğin, müşteri şikayetlerinin daha çok hangi konularda yapıldığı analiz edilirken metin madenciliği tekniklerinden yararlanılır. Gürkan Aydın'ın 2017 yılında yaptığı çalışmada 3 farklı gsm şirketinin (Türk Telekom, Turkcell, Vodafone) şikayetvar.com sitesi üzerindeki şikâyetleri incelenerek bu şikâyetlerin hangi konularda, ne oranda yapıldığı bilgisi elde edilmiştir. Joker yöntemiyle internet, fatura, kapsama ve hizmet kategorileri ile bir sözlük oluşturulmuş ve sözlük sayesinde hangi gsm şirketinin hangi konularda şikâyet aldıkları tespit edilmiştir. Bu şekilde tüketiciler bu grafikleri inceleyip gsm operatörlerini seçebilmektedirler. Firmalarda eksik olduğu konularda iyileştirme sürecine gidebilmektedirler.

2.2.2. Bilimsel ve Medikal İncelemeler

"Yazar tespiti, akademik bir çalışmada intihal tespiti, sağlık ile biyoloji alanındaki en son yenilikler ve mevcut makaleler ile ilgili bilgilerin toplanılarak ve özet haline getirilerek ilgili birimlere özel gruplandırılması, bu yolla her türlü yeniliğin takibi yapılabilir yeni çalışmalar için kaynak oluşturulabilir (Çelikyay, 2010: 49).

2.2.3. Pazar Araştırması

"Yayınlanmış belgeler, basın bültenleri ve web sayfaları pazar etkisinin ölçülmesi için aranır ve izlenir. Metin madenciliği niceleyici yöntemler ile açık

uçlu anket soruları ve mülakatların değerlendirilmesinde kullanılabilmektedir." (Melek, 2012: 29)

2.2.4. Finans

Metin analizi ile finans konulu tüm haberlerin anlık olarak işlenmesi ile birlikte finansal çıkarım, spekülatif haberlerin, satın almaların tespiti yapılabilir. Suat Atan 2016 yılında yaptığı "Metin madenciliği ile sentiment analizi ve borsa istanbul uygulaması " metin madenciliğinin finans alanında kullanıma güzel bir örnektir. Uygulamada, Borsa İstanbul'da, BIST30 endeksi altındaki şirketlerin 2014 yılı içerisinde, finansal durumları ile bu şirketler hakkında çıkan haberler arasındaki ilişkilerin analizini ele alınmaktadır.

2.2.5. Tıp

Tıp alanındaki verilerin çoğu karmaşık, yapılandırılmamış formatta, kağıt tabanlı veya elektronik ortamda serbest metin olarak yer almaktadır. Hekimlerin hasta ile ilgili konularda karar verme sürecinde veya araştırma yaparken hasta raporları, klinik çalışmalar, araştırma raporları, web sayfaları ve hastane kayıtları gibi yapılandırılmamış formattaki veriler kullanılmaktadır. Yapılandırılmamış formatta bulunan bu veri yığınlarını insan gücüyle analiz etmek ve istenilen bilgiye ulaşmak hem zordur hem de zaman kaybına yol açmaktadır. İşte bu aşamada metin madenciliği teknikleri ile istenilen bilgiye ulaşmak daha kolay hale gelmektedir.

Metin madenciliği teknikleri sayesinde tıbbi araştırma sayfaları, makaleler, haberler kullanılarak belirtiler, ilaçlar, hastalıklar, kimyasallar arasındaki önemli ilişkilerin çıkarılması sağlanılmıştır.

"Metin madenciği, tıp alanında özellikle tıbbi araştırmalarda, belirtilerle hastalıklar ve ilaçlarla kimyasal maddeler arasında nedensel ilişkileri belirlemede, hasta kayıtlarının analiz edilmesinde, gen-gen ve protein-protein ilişkilerinin tanımlanmasında, tanı ve tedavileri geliştirmek, servis kalitesini ve faydayı arttırmak, maliyetleri kontrol etmek için kullanılmaktadır." (Oğuz, 2007: 8-9.)

Shatgay' ın çalışmasında, DNA mikroarray deneylerinde genler arasındaki fonksiyonel ilişkileri keşfederken, Swanson' da çalışmasında, metin madenciliği teknikleri ile hastalıklar ve belirtileri arasındaki ilişkilerin ve bağıntıların bulunması sağlamıştır.

2.2.6. İnternet- Yazılım Alanı

Metin madenciliği tekniklerinden yazılım-internet alanında da büyük ölçüde yararlanılmaktadır." Araştırma yapılacak sitelerdeki yasal olmayan ifadelerin tespit edilebilmesi, spam maillerin ayıklatılabilmesi, bir yazılımın planlama safhasındaki dokümantasyonlarına bakılarak yazılımın kaynak kodu ile örtüşüp örtüşmediğinin bulunulabilmesi, bir metnin hangi dilde yazıldığının tespit edilebilmesinde metin madenciliği tekniklerinden yararlanılmaktadır. " (Çelikyay, 2010: 50)

2.3.METİN MADENCİLİĞİ UYGULAMA ADIMLARI

Metin madenciliği 5 adımda gerçekleştirilir. Bunlar;

2.3.1. Metin Topluluğu Oluşturma

Metin topluluğu oluşturma, çalışma yapılacak olan konularda çeşitli sorgu araçları, motorları kullanılarak bir metin madeni oluşturma sürecidir. Özellikle çağımızda bilgiye erişim, kaynak sorgulama en çok arama motorları üzerinden gerçekleştirilmektedir. Bunlar google, yandex, google scholar, opera vb. arama motorlarıdır. "Bu tarz sorgulama araçları ile toplanan veriler dışında veritabanları, kişisel bilgisayarlarda bulunan metin benzeri veriler (fatura, hesap belgesi, mektup, mail, epikriz gibi.) ile de metin madeni oluşturulabilir." (Oğuz, 2009: 10)

2.3.2. Metin Ön İşleme

Metin ön işleme süreci, metin madenciliği çalışmalarının en önemli aşamalarından birisidir. Bu aşamada çalışmada metin içerisinde bize yararı

olmayan, yapacağımız araştırmada doğru sonuçlar çıkarmayan kelimeler ayıklanır. Bu işlem daha güvenilir, anlamlı bilgiye ulaşmamızı sağlar. Metin ön işleme süreci de 4 aşamadan oluşmaktadır.

2.3.2.1. Ayrıştırma

"Metin madenciliğinde, metin ön işleme aşamasında yapılacak ilk işlem, metinlerin analiz edilebilmesi ve sınıflandırma işlemleri için hazır hale getirilmesidir. Bunun için ilk önce metindeki XML ve HTML gibi her türlü etiket ifadelerinin çıkarılması gerekir. Bununla birlikte harf olarak ifade edilemeyen karakter boşluklarla yer değiştirir. Tek harfli sözcükler çıkarılır. Her bir karakter analiz yapılacak dile dönüştürülür." (Çelikyay; 2010: 60)

2.3.2.2. Durdurma Kelimelerinin Çıkarılması

İncelenecek olan metinde çok fazla kullanılan fakat tek başına bir anlam ifade etmeyen edat, bağlaç, harf ve herhangi bir anlam ifade etmeyen kelimelerin metin içerisinden çıkarılmasıdır. Aynı zamanda noktalama işaretleri de ilgili amaç için gerekli değil ise metin içerisinden atılmalıdır. Tek başlarına bir anlamı olmayan kelimeler durdurma kelimeler (stop words) olarak adlandırılmaktadır. Durdurma kelimeler, çalışmanın doğru bir şekilde yapılabilmesi için bir dosya içerisinde saklanmalıdır.

2.3.2.3. Gövdeleme

Bütün temizleme ve ayrıştırma işlemleri yapıldıktan sonra kelimeleri kökleri bulunur. Türkçe sondan eklemeli bir dil olduğu için bir sözcüğün aldığı her ek ona farklı bir anlam kazandırmaktadır. Örneğin, kitap, kitaplık, kitaplar, kitapçı her biri farklı anlam ifade etmesine rağmen kelimelerin kökü kitaptır. Gövdeleme işlemi 2 gruba ayrılmaktadır.

2.3.2.3.1. Gövdeleme

Gövdeleme yöntemi, direkt olarak kelimenin kökü ile ilgilenir. Kelime köklerinin belirlenebilmesi için kelimelerin biçimsel ve anlamsal benzerliklerinin bulunması gerekir.

"Kök bulmada karşılaşılabilecek iki sorun vardır: Birincisi, yapılacak işlemde çok ileri giderek birbirinden anlamca çok farklı kelimelerin aynı anlam grubuna bağlanmasıdır. Bu durumda sistem, konuya uygun olmayan dokümanları da konuyla ilgili şeklinde yorumlayabilir. Diğer bir sorun da, kelimelerin köklerine ulaşılmaya çalışılırken çok az ekin çıkarılması işlemidir. Bu durumda da sistem konuya uygun dokümanları, "uygun olmayan" dokümanlar olarak algılayabilir." (Rouka, 2012: 29)

Tablo 10: Bir kelimenin farklı kullanımları ve kökleri

kullanmak	Kullan
kullanım	Kullan
kullanılmış	Kullan

2.2.3.2. Joker Yöntemi

"Joker yöntemi, gövdeleme yöntemiyle benzerlik göstermektedir. Gövdeleme yönteminde çekim ve yapım eklerinden ayrıştırılan kelimeler, ortak bir köke indirgenirken, joker yönteminde köke indirgeme şartı yoktur. Kökün yanına ek de alabilir." (Ilhan, 2001)

Joker yöntemi, hem işlem sayısını azaltırken hem de sözlükte yer alacak kelime sayısını azaltmaktadır. Örneğin, kömür ve kömürlük kelimelerinin geçtiği bir yazıda kömür* kelimesini joker kelime olarak belirlenir. Bundan sonra kömür kelimesinin aldığı her ek joker kelime olan kömür terimi olarak işlem görecektir. * işareti kelimenin joker kelime olduğunu temsil eder.

2.3.3. Özellik Belirleme

Metin madenciliği uygulamalarında çalışılacak metnin analize uygun hale getirilebilmesi için metnin içinde geçen ilgili kelimelerin keşfedilmesi gereklidir. Özellik belirleme işlemi ise tam bu aşamada karşımıza çıkmaktadır. "Özellik belirleme, ön işlemden geçen metnin içerisinden önemli kelimelerin tespiti ve ilgisiz kelimelerin (sadece birkaç dokümanda gözlemlenen özelliklerin çıkarılması, birçok dokumanda gözlemlenen özellikleri azaltma vb.) çıkarılması işlemidir." (Oğuz, 2009: 9) Özellik belirleme sayesinde iş yükü azaltırken aynı zamanda daha doğru, güvenilir sonuçlar elde edilir, veri kalitesini artırır.

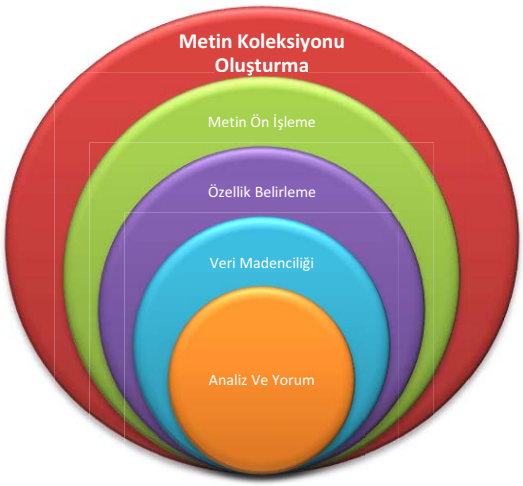
2.3.4. Veri Madenciliği

Metin madenciliği yapılandırılmamış veriler üzerinde çalışırken, veri madenciliği yapılandırılmış veriler üzerinde çalışmaktadır. Metin madenciliğinde yapılandırılmamış veriler, yapılandırılmış hale getirilerek analize için uygun formata dönüştürülmüş olur. Metin madenciliği uygulama adımlarındaki veri madenciliği aşaması, yapılandırılmış bir formata dönüştürülen metinlerin geleneksel veri madenciliği teknikleriyle (karar ağaçları, yapay sinir ağları, kümeleme, genetik algoritma vb.) analizi sürecidir.

2.3.5.Analiz Ve Yorum

Veri madenciliği teknikleri ile elde edilen sonuçların değerlendirilip kullanıcıya uygun, doğru ve anlaşılır şekilde teslim edilmesi aşamasıdır.

Şekil 5: Metin Madenciliği Uygulama Adımları



2.4.METİN GÖSTERİMİ

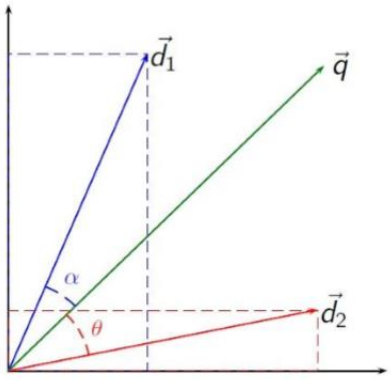
Metinler sayısal halde bulunmadığı için metinler üzerinde sayısal işlemlerin yapılması çok zordur. Bu yüzden metinler farklı gösterim şekillerine dönüştürülür. Bunlardan bir tanesi vektör uzay modelidir.

2.4.1.Vektör Uzay Modeli

Vektör uzay modeli bilgi çıkarımı, bilgi filtreleme, indeksleme gibi alanlarla kullanılan cebirsel bir modeldir. Doğal dil ile yazılmış dokümanların çok boyutlu uzayda temsil edilmesini sağlar. Yapısal olmayan verilerin yapısal hale dönüştürülmesinde kullanılır.

Şekil 6: Vektör Uzay Modeli

d1=2w1 + 3w2 +1w3 , d2=5w1 + 2w2 +3w3 , q=7w1 + 3w2 +5w3



"Q vektörü, sınıflandırılacak doküman, D1 ve D2 vektörleri ile kıyaslanır ve vektörler arasındaki açının kosinüsü hesaplanarak kıyaslama yapılır. Burada w' ler kelimelerin frekanslarını ifade etmektedir. Bu kıyaslamada kosinüs değeri hangisinde yüksek olursa Q vektörü o dokümanın sınıfına dahil edilir. Örnek olarak Q vektörü ile D1 vektörü arasındaki açının kosinüsü, Q vektörü ile D2 vektörü arasındaki açının kosinüsünden daha büyükse, o zaman, Q vektörü ile D1 vektörü ile aynı sınıfa dahil edilir. Yani: Eğer Cos(D1 , Q) > cos(D2 , Q) → Q vektörü D1 le aynı kategoridedir." (Leila ROUKA,2012)

Vektör uzay modeli MxN boyutlu bir uzaydır. Burada M'ler tüm dokümandaki her bir farklı kelime sayısını ifade ederken N'ler ise elimizdeki dokümanların tümünü temsil eder. Vektördeki her eleman değeri, kelimenin metinde ne sıklıkla geçtiğini sayısal olarak temsil etmektedir. Buna kelimelerin ağırlıklandırılması denilmektedir.

2.4.2. Kelime Filtreleme Ve Kelime Ağırlıklandırma

İsmail Ferhat Pilavcılar, 2007 yılında yaptığı çalışmada kelimelerin ağırlıklandırılması konusunu bitsel ve frekans olarak 2 şekilde açıklamıştır.

Kategorisini bulmak istediğimiz metin aşağıdaki şekilde olursa;

SORGU: Taraftarlar hakeme tepki gösterdiler. Hakem sahayı terk etti.

Örnek eğitim dokümanları aşağıdaki gibi olursa;

1-Gribe yakalanan hasta grip olduğunu anlamamıştı. İlacını almamıştı. (Sağlık)

2-İlacını aksatanlar hastalığa davetiye çıkarırlar.
(Sağlık)

3-Yıllık enflasyon oranı bu senede yükselişte. (Ekonomi)

4-Tarımla uğraşanlar bu yıl tarımdan zarar edecekler. (Ekonomi)

5-Hakemin gözü önünde olmasına rağmen hakem penaltı çalmadı. (Spor)

6-Taraftarlara erken gelen gol ilaç gibi geldi ve taraftarlar golden sonra hiç susmadı. (Spor)

Tablo 11: Kelime Filtreleme

KELİME	JOKER
Enflasyon*	
Grip*	Grib*
Hakem*	
İlaç*	İlac*
Taraftar*	
Tarım*	

Sözlük={ enflasyon*, grip* , hakem*,İlaç*, taraftar*,tarım*}

Eğer vektörlerimizi kelime frekansına göre ifade edecek olursak sözlükte yer alan kelime ilgili dokümanda kaç kere geçiyorsa yazılır.

D1=(0,2,0,1,0,0)

D2=(0,0,0,1,0,0)

D3=(1,0,0,0,0,0)

D4=(0,0,0,0,0,2)

D5=(0,0,2,0,0,0)

D6=(0,0,0,1,2,0)

DQ=(0,0,2,0,1,0) ➔ Sorgu vektörü

Eğer vektörlerimizi bitsel olarak ifade etmek istersek sözlükteki kelime ilgili dokümanda geçiyorsa 1 geçmiyorsa 0 değerini alır.

D1=(0,1,0,1,0,0)

D2=(0,0,0,1,0,0)

D3=(1,0,0,0,0,0)

D4=(0,0,0,0,0,1)

D5=(0,0,1,0,0,0)

D6=(0,0,0,1,1,0)

DQ=(0,0,1,0,1,0)

2.5.BİRLİKTELİK KURALLARI

Birliktelik analizi, büyük veri kümesi içerisinde geçmişte sıkça birlikte ortaya çıkan özelliklerden yola çıkarak geleceğe yönelik tahminlerin yapılmasını sağlayan denetimsiz bir öğrenme biçimidir. Birliktelik analizi ile aynı anda ortaya çıkan olaylar incelenir. Birliktelik analizine pazar sepeti analizi de denilmektedir. Böylece müşterilerin satın alma alışkanlıkları tespit edilir. Örneğin; kahve alan kişiler %70 olasılıkla kahvenin yanında çikolata da almaktadır. İşletmeciler bu sonucu göz önüne aldığında kahve standının yanına çikolata standı da koymaktadır.

"Birliktelik algoritmaları basit kategorik değişkenler, ikili (dichotomous) değişkenler ve/ve ya çoklu hedef değişkenleri analiz etmede kullanılmaktadır. Algoritma veride ya da herhangi birliktelikte (ön tanımlı varyant hariç) mevcut farklı kategorilerin sayısını belirtmenizi gerektirmeden birliktelik kurallarını belirleyecektir. Bir çapraz çizelgeleme (corss-tabulation) biçimi değişkenler ya da kategorilerin belirlenmiş sayılarına ihtiyaç duymadan yapılandırılabilir." (Melek, 2012: 40)

Başak Oğuz 2009 yılında yapmış olduğu çalışmada birliktelik kuralını şu şekilde tanımlamıştır;

$A_1, A_2, ..., A_m \quad => \quad B_1, B_2, ..., B_n$

Bu ifadede yer alan, A_m ve B_n, yapılan iş veya nesnelerdir. $A_1, A_2, ..., A_m$ eylemleri gerçekleştiği zaman, sık bir şekilde $B_1, B_2, ..., B_n$ eylemleri de aynı olay içinde meydana geldiğini göstermektedir.

Birliktelik kuralında iki önemli kriter vardır. Birincisi destek, ikincisi güven kriteridir. Birliktelik kuralı bu iki kriter ile hesaplanmaktadır. Destek kriteri, öğeler arasındaki ilişkinin ne kadar sık olduğunu, güven kriteri ise A öğesinin hangi olasılıkla B öğesi ile birlikte gerçekleşeceğini ifade eder. İki öğenin birlikteliğinin önemli olması için hem destek, hem de güven kriterinin olabildiğince yüksek olması gerekmektedir. Aşağıdaki formülde A ve B öğelerinin birliktelik kuralı A = > B şeklinde ifade edilmiştir.

Destek (A=>B)= (A ve B'nin bulunduğu satır sayısı) / (toplam satır sayısı)

A=>B birliktelik kuralının güven değeri ise, A olayı gerçekleşince B olayının da gerçekleşmesi yüzdesidir. Örneğin, bir kural % 60 güvenirliğe sahip ise, A'yı içeren ürün setleri % 60'ı B ürününü de içermektedir.

Güven (A=>B) = (A ve B'nin bulunduğu satır sayısı) / (A'nın bulunduğu satır sayısı)

Güven değerinin % 100 olması durumunda, kural bütün veri analizlerinde doğrudur ve bu kurallara "kesin" denir.

Bir başka değer ise lift değeridir. Lift değeri, birliktelik kuralının kalitesini belirler. Güven değerinin, güven değerinin beklenen değerine bölünmesi ile elde edilen ve 0 ile sonsuz arasında yer alan bir değerdir.

Beklenen Güven (A=>B) = Destek (A) * Destek (B) / Destek (A)

Lift (A=>B) = Güven (A=>B) / Beklenen Güven (A=>B)

"Lift değeri birden küçük ise A olayının meydana gelmesinin B olayı üzerinde negatif bir etki yaratırken, lift değeri 1'e yakın ise A olayının meydana gelmesinin B olayı üzerinde hiçbir etkisi olmadığını, yani birbirinden bağımsız olaylar olduğunu göstermektedir. 1'den büyük olması ise A olayının meydana gelmesinin B olayı üzerinde pozitif bir etkisi olduğunu belirtmektedir. Lift değerinin 1'den büyük olması kuralın kalitesinin iyi olduğunu göstermektedir." (Oğuz,2009: 7)

Tablo 12: Birliktelik Kuralı

TABLE 7.1 Word Correlations, with Their Support and Confidence Values. Support Is Expressed by the Joint Probability of Word 1 and Word 2 Occurring Together; Confidence Is the Conditional Probability of Word 1 Given Word 2.

Summary of association rules (Scene 1.sta) Min. support = 5.0%, Min. confidence = 5.0%, Min. correlation = 5.0% Max. size of body = 10, Max. size of head = 10						
	Body	==>	Head	Support(%)	Confidence(%)	Correlation(%)
154	and, that	==>	like	6.94444	83.3333	91.28709
126	like	==>	and, that	6.94444	100.0000	91.28709
163	and, PAROLLES	==>	will	5.55556	80.0000	73.02967
148	will	==>	and, PAROLLES	5.55556	66.6667	73.02967
155	and, you	==>	your	5.55556	80.0000	67.61234
122	your	==>	and, virginity	5.55556	57.1429	67.61234
164	and, virginity	==>	your	5.55556	80.0000	67.61234
121	your	==>	and, you	5.55556	57.1429	67.61234
73	that	==>	like	6.94444	41.6667	64.54972
75	that	==>	and, like	6.94444	41.6667	64.54972
161	and, like	==>	that	6.94444	100.0000	64.54972

Kaynak: Nisbet ve diğerleri, 2009: 127.

Yukarıda örnek olarak verilen tabloda değişkenler arası destek, güvenirlilik ve korelasyon değerleri ifade edilmektedir. Çok büyük veri setlerinin analizi yapılırken bu çapraz tablolama tekniğinden yararlanılabilir. Destek (support) değeri iki değişkenin birlikte ortaya çıkma olasılığını gösterir. Tablo yer alan destek değeri ile metin içerisindeki "and, your, virginity" kelimelerinin birlikte ortaya çıkma olasılığı %5.55 olarak ifade edilmektedir. Ayrıca "and, that ve like" kelimeleri arasındaki ilişki de %91 ifade edilmektedir. "and, that ve like" kelimeleri arasındaki ilişki oldukça yüksektir.

Güvenirlilik (Confidence) değeri ile bir durum meydana geldiğinde diğer bir durumun da birlikte ortaya çıkma olasılığını ifade edilmektedir. Yukarıdaki tablodan da görülebileceği gibi ilk satırda "will" kelimesinin geçtiği bir cümlede "and, PAROLLES" kelimelerinin de geçme olasılığı %66.6 olarak ifade edilmektedir. Tekrar "and, that ve like" kelimeleri ele alındığında "and, that" kelimelerinin geçtiği bir cümlede "like" kelimesinin de geçme olasılığı %83.33 olarak ifade edilmektedir.

2.6.KÜMELEME ANALİZİNDE KULLANILAN BENZERLİK VE FARKLILIK ÖLÇÜLERİ

2.6.1.Öklit Uzaklık Ölçüsü

Birey veya nesneler arasındaki uzaklıkları hesaplamak için en çok kullanılan uzaklık ölçüsü Öklid uzaklık ölçüsüdür. "Öklid uzaklık ölçüsü 2 nesne arasına çizilecek bir düz doğrunun uzunluğunu esas alır. Öklid uzaklıkları kümeleme analizine sıra dışı olabilecek yeni nesnelerin eklenmesinden etkilenmezler. Ancak boyutlar arasındaki ölçek farklılıkları Öklid uzaklıklarını önemli ölçüde etki etmektedir. "(Demiralay, 2005)

$$d(i,j) = \sqrt{(x_{i1}\text{-}x_{j1})\text{^}2 + (x_{i2} - x_{j2})\text{^}2 + \cdots + (x_{ip} - x_{jp})\text{^}2}$$

Burada d(i,j) i ve j biriminin birbirine olan uzaklığını gösterir.

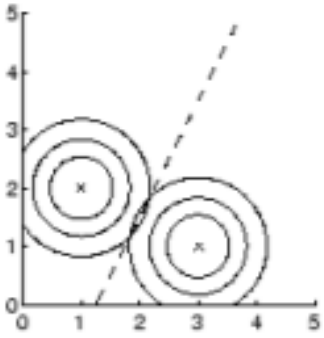
n=Birim sayısı

p=Değişken sayısı

i,j= 1,2,3,…,n

"Nesnelerin konumları ele alınarak birbirlerinden ne ölçüde farklı oldukları tespit edilir. Veri seti yoğun ve birbirlerinden iyi ayrılmış kümeler içeriyor ise iyi sonuç elde edilir. İki nesne birbirine ne kadar yakın ise Öklid uzaklığı da o kadar sıfıra yaklaşır. Aşağıdaki şekilde de görüldüğü gibi Öklid uzaklığı kullanılarak bulunan kümeler küresel bir yapıya sahiptir." (Işık, 2008: 38)

Şekil 7: Öklid uzaklığının kümeleme özelliği



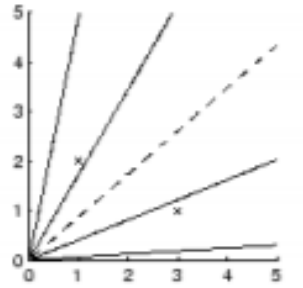
2.6.2. Kosinüs Benzerliği

Metin kümelemede en sık kullanılan vektör tabanlı bir ölçüt olan Kosinüs Benzerliği ile, iki vektör arasındaki açının kosinüs değeri hesaplanarak vektörlerin benzerliği elde edilir. Metin belgelerinde, çok sayıda bilgi çıkarımında ve kümeleme de kullanılır. Kosinüs benzerliğinin en önemli özelliği, belge uzunluğunun bağımsız olmasıdır. 2 belge arasındaki kosinüs benzerlik değeri 1 ise, bu iki belge özdeştir yani birbirlerine benzerdir.

$$\cos(\theta) = \frac{d \bullet d^*}{|d||d^*|} = \frac{\sum_{i=1}^{n} d_i d_i^*}{\sqrt{\sum_{i=1}^{n} (d_i)^2} \sqrt{\sum_{i=1}^{n} (d_i^*)^2}}$$

"Burada d ve d* birbirinden farklı iki belgeyi temsil eden çok boyutlu vektörleri ve "•" vektörlerin iç çarpımını, $|d|$ ise vektörün uzunluğunu temsil etmektedir." (Meltem Işık, 2008: 38)

Şekil 8: Kosinüs benzerliğinin kümeleme özelliği



2.6.3. Pearson Uzaklık Ölçüsü

"Pearson ilişkisinde, Kosinüs benzerliğindeki gibi iki vektörün benzerliğinde vektörlerin aralarındaki açıya göre karşılaştırma yapılır. Kosinüs benzerliğinden farkı iki vektörün iç çarpımı yapılmadan önce her birinin ayrı ayrı ortalama değerleri hesaplanır ve her ortalama değer ait olduğu vektörün tüm elemanlarından çıkarılır. d ve d* birbirinden farklı iki belgeyi temsil eden çok boyutlu vektör ise aralarındaki Pearson benzerliği aşağıdaki gibi hesaplanır.

$$S^{Pearson}(d,d^*) = \frac{1}{2}\left(\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}_i\right)\left(d_i^* - \bar{d}_i^*\right)}{\sqrt{\sum_{i=1}^{n}\left(d_i - \bar{d}_i\right)^2 \sum_{i=1}^{n}\left(d_i^* - \bar{d}_i^*\right)^2}} + 1 \right)$$

Çıkan sonucu [0,1] aralığında tutmak için normalizasyon işlemi yapılmaktadır. Bunun için hesaplanan değere 1 eklenip daha sonra da 2'ye bölmek yeterli olacaktır. Pearson ilişkisi değeri 1'e yaklaştıkça iki vektörün birbirlerine olan benzerlikleri de artar." (Işık, 2008: 39)

2.6.4. Manhattan Uzaklık Ölçüsü

Manhattan uzaklık ölçüsünde birimler arasındaki mutlak uzaklık kullanılmaktadır. Manhattan uzaklığı, birimlerin aynı değişkenleri arasındaki mutlak farkların toplanması olarak ifade edilir ve aşağıdaki formül yardımıyla hesaplanır.

$$d(i,j) = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|$$

Manhattan uzaklığı değişkenler arasında ilişki olmadığını varsaymaktadır. Eğer değişkenler arasında korelasyon (ilişki) varsa Manhattan uzaklık sonuçları anlamlı olmayacaktır

2.6.5. Minkowski Uzaklık Ölçüsü

"Minkowski uzaklık ölçüsü formülünde yer alan m değerinin alacağı farklı değerlere göre yeni formüller türetir. Minkowski uzaklık ölçüsü kullanılarak iki birim arasındaki uzaklık aşağıdaki formül yardımıyla hesaplanır.

$$d(i,j) = \left[\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{m} \right]^{1/m}$$

Minkowski uzaklık ölçüsündeki m değeri büyük ve küçük farklara verilen ağırlığı değiştirir. m=1 değerini alırsa, formül, Manhattan uzaklık ölçüsünün formülüne, m = 2 değerini alırsa, formül Öklid uzaklık ölçüsü formülüne dönüşür." (Dinler, 2004: 15)

2.7.KÜMELEME ANALİZİ

Kümeleme analizi, özellikleri belirlenmemiş, herhangi bir yöntem ile gruplanmamış verileri aralarındaki benzerliklerine göre sınıflandırmayı sağlayan bir yöntemdir. Elde edilen kümeler küme içi homojenlik ve kümeler arası heterojenlik gösterirler. Kümeleme analizindeki amaç elde edilen sınıflardaki verileri yapılan araştırmaya uygun olarak yararlı, özetleyici bilgiler sunmaktır. Eğer yapılan kümeleme işlemi başarılı bir şekilde yapılırsa, bir geometrik çizim yapıldığında nesneler küme içerisinde birbirine çok yakın, kümeler ise birbirinden uzak olacaktır. Kümeler arasındaki benzerlik ve farklılıkların bulunması uzaklık ve yakınlık ölçüleri ile yapılmaktadır.

Kümeleme analizi, veri indirgeme, hipotez oluşturma, doğru modelin belirlenmesi gibi amaçlarla da kullanılmaktadır. Örneğin, Türkiye'deki 81 ilin bulunduğu bir veri kümesini bölge bölge ayırmak ya da baş harfi "İ" ile başlayan iller diye sınıflandırmak bir kümeleme analizidir. Burada yapılan işlem veri indirgemeye bir örnek olacaktır.

Kümeleme analizi 4 aşamadan meydana gelmektedir.

İlk aşaması veri matrisinin oluşturulması aşamasıdır. "Yani ilk olarak doğal sınıflamaları hakkında kesin bilgilerin bulunmadığı ana kütlelerden alınan n sayıda birimin incelenen p sayıda değişkene ilişkin gözlem sonucu değerleri elde edilir." (Yılmaz ve Patır, 2010: 100)

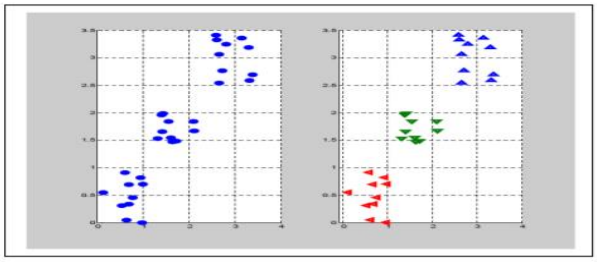
İkinci aşama uzaklık matrisinin elde edilme aşamasıdır. Uzaklık matrisi elde olan verilerin birbirleri arasındaki uzaklık ve benzerlik ölçülerine göre yapılmaktadır. Uzaklık ve benzerlik ölçüleri bir sonraki konu olan kümeleme analizinde kullanılan benzerlik ve farklılık ölçülerinde anlatılmıştır.

Üçüncü aşama tercih edilecek uygun hiyerarşik veya hiyerarşik olmayan kümeleme yöntemlerinden biriyle benzerlik/farklılık matrisindeki birimler / değişkenler kümelere ayrılır.

Dördünce aşama ise yapılan kümeleme analizinin başarısının test edilmesidir. Yani kümelerin kendi içinde ve kümeler arasında ne kadar doğru bir şekilde sınıflandırıldığının analizidir. Daha öncede bahsedildiği gibi eğer kümeleme analizi başarılı bir şekilde yapılmış ise geometrik çıktısında veriler

küme içerisinde birbirine çok yakın, kümeler arasında ise birbirinden çok uzak olacaktır. Aşağıdaki şekilde geometrik gösterimi yapılmıştır.

Şekil 9: Kümeleme Analizi

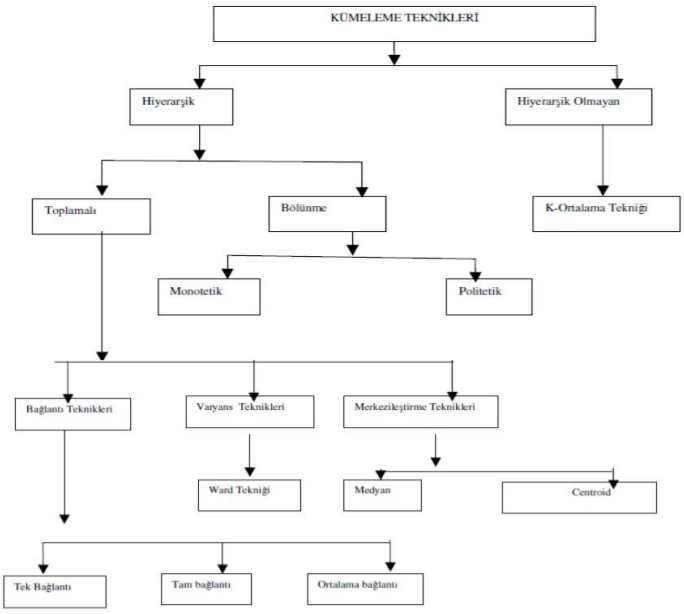


Kaynak: VATANSEVER, Metin (2008), *Görsel Veri Madenciliği Tekniklerinin Kümeleme Analizlerinde Kullanımı ve Uygulanması*, Yayınlanmamış Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, Fen Bilimler Enstitüsü, İstanbul, s. 85.

2.7.1. Kümeleme Yöntemleri

Kümeleme analizi yöntemleri hiyerarşik ve hiyerarşik olmayan kümeleme yöntemleri olmak üzere 2 bölüme ayrılmaktadır. Aşağıdaki akışta kümeleme tekniklerinin kendi içlerinde de nasıl ayrıldıklarını detaylı bir şekilde gösterilmektedir.

Tablo 13: Kümeleme Teknikleri



Kaynak: Akın, 2008: 9

2.7.1.1. Hiyerarşik Kümeleme Yöntemi

"Hiyerarşik kümeleme yöntemleri, veri setinin birimlerinin birbirlerine olan uzaklık değerlerini kullanarak, veri setindeki birimlerin hiyerarşik ayrıştırmasını yapar. Hiyerarşik ayrıştırma sırasında, dendogram olarak bilinen ağaç diyagramı kullanılır. Ağaç diyagramı, hiyerarşik kümeleme yöntemiyle elde edilen kümelerin görselleştirilmesini sağlar. Küme sayısına görsel olarak karar verilir." (Dinler, 2014: 17)

2.7.1.2. Hiyerarşik Olmayan Kümeleme Yöntemi

"Hiyerarşik olmayan kümeleme yöntemleri ise n adet birimden oluşan veri setini başlangıçta belirlenen k<n olmak üzere k adet kümeye ayırmak için kullanılmaktadır. Bu özellik hiyerarşik olmayan kümeleme yöntemlerinin, hiyerarşik kümeleme yöntemlerinden ayıran özelliklerden biridir. Hiyerarşik olmayan kümeleme yöntemleri, hiyerarşik kümeleme yöntemlerine oranla daha büyük veri setlerine uygulanabilmektedir. Hiyerarşik olmayan kümeleme yöntemleriyle oluşturulacak k adet kümede her bir küme en azından bir birim

içermekte ve her birim yalnızca bir gurupta bulunmaktadır. “ (Dinler, 2014: 22-23)

Hiyerarşik kümeleme yöntemleri arasında en bilinenleri K-ortalamalar kümeleme yöntemidir.

2.7.1.2.1. K-Ortalamalar Yöntemi

K-ortalamalar yöntemi bilinen en eski kümeleme yöntemlerinden biridir. 1967 yılında J.B. MacQueen tarafından geliştirilmiştir. K ortalamalar yöntemi her verinin yalnızca bir kümeye ait olduğunu belirtmektedir. K – ortalamalar yönteminde, elde bulunan n tane nesne k tane kümeye dağıtılmaktadır. Burada en önemli nokta k küme sayısının önceden biliniyor olmasıdır ve yine en önemli nokta k küme sayısının belirlenmesidir.

“K- ortalamalar yönteminde yapılacak işlemler aşağıdaki gibidir.

1. K adet birim başlangıç küme merkezleri olarak rastgele seçilir.

2. Küme merkezi olmayan birimler belirlenen uzaklık ölçütlerine başlangıç küme merkezlerinin ait oldukları kümelere atanır.

3. Yeni küme merkezleri oluşturulan k adet başlangıç kümesindeki değişkenlerin ortalamaları alınarak oluşturulur.

4. Birimler en yakın oldukları oluşturulan yeni küme merkezlerine birimlerin uzaklıkları hesaplanarak kümeye atanır.

5. Bir önceki küme merkezlerine olan uzaklıklar ile yeni oluşturulan küme merkezlerine olan uzaklıklar karşılaştırılır.

6. Uzaklıklar makul görülebilir oranda azalmış ise 4. adıma dönülür.

7. “Eğer çok büyük bir değişiklik söz konusu olmamış ise, tekrarlama sona erdirilir.” (Dinler, 2014: 23)

K ortalamalar yönteminin amacı, hata kareler toplamını minimize edecek olan bölünme yani k sayısını bulmaktır.

K değerinin belirlenmesi oldukça önemlidir. Özellikle veri sayısının çok olduğu kümelerde k sayısının belirlenmesi zordur.

Mehmet Dinler 2014 yılında yapmış olduğu çalışmasında k sayısının belirlenmesinde

$$k = \sqrt{\frac{n}{2}}$$

formülünü kullanmıştır. Bu ilk yapılan çalışmalardan en çok bilinen yöntemdir. Fakat veri sayısının çok olduğu gözlemlerde doğru sonuçlar alınmayabilmektedir.

k= küme sayısı

n= birim sayısı

Bir başka yöntem ise 1971 yılında Marriot tarafından grup içi kareler toplamı matrisinden yararlanılarak bulunmuştur.

$$M = k^2|W|$$

Burada W grup içi kareler toplamı matrisini ifade etmektedir.

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \overline{x_j})(x_{ij} - \overline{x_j})'$$

nj; j. kümedeki birim sayısı

k ; küme sayısı

xij; j. kümedeki i. birim değerleri

jx ; j. kümenin örneklem ortalama vektörü

2.8.TEMEL BİLEŞENLER ANALİZİ

"Temel bileşenler analizi, veri tanıma, sınıflandırma, görüntü sıkıştırma alanlarında kullanılan, bir değişkenler setinin varyans - kovaryans yapısını, bu değişkenlerin doğrusal birleşimleri vasıtasıyla açıklayarak, boyut indirgenmesi ve yorumlanmasını sağlayan, çok değişkenli bir istatistiksel yöntemdir." (Yıldız, Doğan ve Çamurcu, 2010 :210)

Temel bileşenler analizinin 2 amacı vardır. Bunlar p tane değişken; doğrusal, ortogonal ve birbirinden bağımsız olma özelliklerini taşıyan k tane (k<=p) yeni değişkenle açıklanması ve boyut indirgeme amacıdır.

Veri sayısının çok olduğu dilbilimsel girdilerde veri sayısı azaltılarak hem zaman hem de hata sayısını azaltmada oldukça etkilidir.

Boyut indirgeme aşamasında dikkat edilmesi gereken nokta, yararlı bilgilerin de silinmemesi için azaltılan değişken ve bilgi kaybı arasında kabul edilebilir bir değişimin olmasıdır.

Temel bileşenler analizi 4 aşamada gerçekleştirilir.

1-) Gerekli olan durumlarda veriler ortalama yöntemiyle düzgünleştirilir.

2-) Kovaryans matrisi hesaplanır.

3-) Özdeğer - Özvektör değerleri hesaplanır.

4-) İndirgeme için özellik vektörü seçilir ve indirgeme işlemi yapılır.

p tane değişken; doğrusal, ortogonal ve birbirinden bağımsız olma özelliklerini taşıyan k tane (burada k<=p olmalıdır.) yeni değişkenle açıklanması temel bileşenlerin boyut indirgeme amacını oluşturur.

Temel bileşenler analizi ve K-ortalamalar yöntemi ile ilgili uygulamalardan birine 2016 yılında Nilgün Şengöz ve Gültekin Özdemir tarafından yapılan "Temel Bileşenler Analizi ve K-ortalama Kümeleme Yönteminin Birlikte Kullanımı: Bir Örnek Uygulama" adlı makale örnek gösterilebilmektedir.

Yapılan çalışma da üç farklı buğday türüne (Kama, Rosa ve Canadian) ait geometrik şekillerine göre yedi kategorili (kernel buğdayının alanı, çevresi, uzunluğu, genişliği, oyuğu, yoğunluğu ve asimetrik katsayısı), her bir elementten 69 adet bulunan 207 tane örnek seti ele alınmaktadır. İlk basamakta temel bileşenler analizi kullanarak varyansların maksimizasyonu sağlanılmış daha sonra k-ortalama yöntemi ile veri setleri kümelenmiştir. K-ortalama yönteminde 3 buğday çeşidi olduğu için küme sayısı 3 alınmaktadır.

K-ortalama yönteminde ilk işlemde her bir kümede 69 gözlem olmasına rağmen kümeleme yöntemi ile birlikte sırasıyla kümelerdeki gözlem sayısı, 1. kümede 61, ikinci kümede 74 ve 3. kümede ise 72 gözlem olmuştur.. Sonuç olarak, Rosa kümesi için 8 adet gözlem, Canadian kümesi için 5 adet gözlem ve Kama kümesi için ise 3 adet gözlem doğru hesaplanamamıştır. Toplamda ise 16 adet gözlem doğru kümelenememiştir. Temel bileşenler analizindeki temel amaç, aralarındaki korelasyonu en az olan birbirlerine dik, veri setinin büyük çoğunluğunu kapsayacak yani onu ifade edecek iki bileşeni bulmaktır. Bu neden ile en yüksek özvektör değerine sahip olan bileşenler bulunulmuş ve bu bileşenlerin toplam veri setini açıklama yüzdeleri hesaplanılmıştır. Çıkan sonuçta bu 2 bileşende bulunan 7 kategoriden 2 tanesi en yüksek değeri almıştır. Bunlar 1.bileşende alan kategorisi, 2. bileşende asimetrik katsayı kategorisi olmuştur. Bu iki kategori birbirlerine dik ve aralarında korelasyon yok ya da çok az korelasyon

olduğu söylenilebilmektedir. Böylece 7 kategoriden 5 tanesi elenmiş ve 2 kategori ile tekrar k-ortalama yöntemi uygulanılmış ve 1. Küme 62, 2.küme 71 ve 3. Küme 74 elemandan oluşmaktadır. Uygulamanın başında küme eleman sayıları sırasıyla 69, 69, 69 idi. Daha sonra 61,74 ve 72 elemandan kümelenmiştir. Fakat k-ortalama yöntemi ile temel bileşenler analizi yöntemi birlikte uygulandığında ilk başta yanlış kümelenen 16 eleman yanlış kümelenirken daha sonra bu sayı 13 elemana düşürülmüştür. Bu sonuç bize k-ortalama yöntemi ile temel bileşenler analizinin birlikte uygulanması ile daha doğru sonuçların elde edilebileceğini göstermektedir.

2.9.DUYGU ANALİZİ(SENTİMENT ANALYSİS)

Metin madenciliği çalışma alanlarından biri duygu analizidir. Duygu analizi, incelenen bir metinden duygu çıkarımı yapmaktadır. Buradaki amaç metni yazan kişinin görüşünü tespit etmektir. Örneğin; bir tatil sitesindeki bir otel ile ilgili yorumlardan yola çıkarak duygu analizi yöntemler ile yorumu yazan kişinin otel ile ilgili görüşleri tespit edilebilir.

"Duygu analizi çalışmaları doğal dil işleme, makine öğrenmesi, hesaplamalı dilbilim, sembolik teknikler gibi yaklaşımları kullanır. Makine öğrenmesi, verilen bir problemi ortamdan edindiği bilgiye göre modelleyen Yapay Zekâ disiplininin bir alt dalıdır. Makine öğrenmesi teknikleri denetimli ve denetimsiz öğrenme metotlarından oluşur. Denetimli öğrenme, önceden gözlemlenmiş ve sonuçları bilinen verileri kullanarak bu verileri ve sonuçlarını ele alan bir fonksiyon oluşturmayı amaçlayan makine öğrenimi tekniğidir. Denetimsiz öğrenme, sonuçları bilinmeyen verideki gizli yapıyı açığa çıkarma tekniğidir. Yani, veriler arasında var olan ama gözle görülmeyen bağıntının açığa çıkarılması işlemidir. " (Nizam ve Akın, 2014: 2)

Duygu analizi yapılırken, ilk aşamada incelenen metinler pozitif, negatif ve nötr olmak üzere sınıflandırılır. Daha sonra görüş tarafı belli olan (pozitif / negatif / nötr) metinsel veriler kullanılarak bir eğitim veri seti oluşturulur ve bu veriler metin madenciliği ön işleme yöntemleri olan iletideki simge ve noktalama işaretlerinin temizlenmesi, kelimeleri köklerine ayırarak terim listelerinin oluşturulması, metin içerisindeki edat, bağlaç ve zamirlerden oluşan durak

kelimelerin kaldırılması, terim frekansları ve ters doküman frekansları yardımıyla vektör uzay modelinin oluşturulması sayılabilir. Vektör uzay modellerinde binary, terim frekansı, ters doküman frekansı, n gram gibi çeşitli yöntemler kullanılmaktadır. Vektör uzay modeli oluşturulan bu veriler, makine öğrenmesinde sıklıkla kullanılan yapay sinir ağları, destek vektör makinaları, karar ağaçları gibi sınıflama algoritmaları yardımıyla eğitilmektedir. Çeşitli sınıflayıcılar tarafından eğitilen modeller yardımıyla yeni gelen duygu ve görüşler sınıflandırılmaktadır.

Örnek olarak sosyal medya üzerinden bir otel ile ilgili yazılmış yorumlardan oluşan bir veri seti oluşturulmaktadır. Daha sonra bu veri seti manuel olarak metinleri pozitif, negatif, nötr olmak üzere 3 veri setine ayrılmaktadır. Bu ayırma işlemini daha önce belirlenen eğitim veri setine göre yapılmaktadır. Metinler, metin madenciliği ön işleme aşamalarından geçtikten sonra çeşitli sınıflandırma algoritmaları (Naive Bayes (NB), Random Forest (RF), Sequential Minimal Optimization (SMO), Decision Tree (J48), 1-Nearest Neighbors (IB1),…) yardımıyla sınıflara ayrılmaktadır. Bu sınıflandırma algoritmalarının başarılarının test edilmesi için model başarım ölçütleri **(** Doğruluk–Hata Oranı (Accuracy-Error Rate), kesinlik (Precision), duyarlılık (Recall), F-Ölçütü (F-Measure)) ve Kappa istatistiği kullanılmaktadır.

Tablo 14: Model Başarım Ölçütlerinin Karşılaştırılması

Sınıflandırma	Doğruluk-hata oranı	Kesinlik	F-ölçütü	Duyarlılık	Kappa istatistiği
NB	59.30	0.5	0.59	0.58	0.28
RF	45.15	0.78	0.66	0.56	0.21
SMO	78.58	0.85	0.70	0.65	0.85
J48	45.36	0.46	0.53	0.45	0.15
Nearest Neighbors	75.62	0.35	0.55	0.50	0.08

Kappa istatistiği 1 değerine yaklaştıkça gözlenen uyumun şans eseri gerçekleşmediğini ifade etmektedir. En iyi performansı gösteren sınıflandırma algoritması SMO'dur ve %81.36 doğruluk başarı oranı göstermiştir.

2.10.YAZAR TANIMA

Metin madenciliği uygulama alanlarından bir diğeri de yazar tanıma, cinsiyet belirleme ve metnin türünü belirleme çalışmalarıdır. Her yazarın kendine has olan yazma üslubu bulunmaktadır. Kimi noktalama işaretlerini bol kullanırken kimi de sürekli olarak hitap şeklinde yazılarını kaleme almaktadır. Her yazıda okuduğumuz yazarlar diline hakim isek bir yayında yazar belirtilmese de, o eseri kimin yazdığını tahmin edebiliriz. Metin madenciliği burada yazar tanıma araştırmalarının insan faktörü ile değil de makine dili ile yazarların üsluplarını matematiksel dile çevrilmesini sağlar. " Yazar Tanıma, konudan bağımsız olarak elimizde bulunan dokümanların hangi yazar tarafından yazıldığını tahmin etme işlemidir. Yazar tanıma işlemlerinden yararlanılarak metinlerin cinsiyete göre sınıflandırılması çalışması da yapılabilmektedir. Karakter kullanım sıklık değerleri dil içinde birbirinden farklı oldukları için belirli bir konuda yazılmış dokümanların karakter sıklıkları birbirine daha yakın iken farklı konularda yazılmış dokümanların karakter sıklıkları ise birbirinden daha uzaktır." (Doğan, 2006: 1)

Literatürdeki örneklerine bakacak olursak, Diri ve Amasyalı' nın 2006 yılında yapmış olduğu *" Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi"* çalışması güzel bir örnek oluşturur. Çalışmanın amacı, yazarı bilinmeyen bir dokümanın önceden belirlenmiş ve yazarlık özellikleri çıkarılmış 18 farklı yazardan hangisine ait olduğunu tespit etmeye dayanmaktadır. Veri setleri İnternet'te belirli gazete sayfasından (www.hurriyet.com.tr, www.vatanim.com.tr, www.sabah.com.tr) farklı yazarların politika, güncel, spor ve magazin gibi konularda yazıları toplanarak oluşturulmuştur. Eldeki mevcut veri seti 18 farklı yazarın her birine ait 35 farklı yazısından oluşan, 630 adet dokümandan meydana gelmektedir. Yazarı bulunması istenen dokümanlardan istatistiksel veriler, kelime sözlüğünün zenginliği, sık geçen Türkçe kelimeler, dilbilgisi özellikleri, Ngram'lar ve bunların çeşitli birleşimleri kullanılarak on farklı özellik vektörü elde edilmiştir. Daha sonra NB, SVM, C 4.5 ve RF sınıflandırma yöntemleri kullanılarak her bir özellik vektörünün sınıflandırmadaki başarıları karşılaştırılmıştır. En başarılı sonuç SVM ile sınıflandırılmasında %92,5 ile elde edilmiştir. Kullanılan sınıflandırıcıların en başarılısı, orijinal vektörlerle

çalışıldığında açık farkla SVM, özellik azaltılmış vektörlerle çalışıldığında ise Naive Bayes ve SVM'dir. Bu çalışmada dokümanların sınıflandırılmasında, n-gram'ların kullanılmasının yazarlık özelliklerine göre daha başarılı olduğu gözlemlenmiştir.

Tablo 15: Özellik Azaltılmadan Kullanılan 5 Vektörün Sınıflandırma Başarısı

	(%)Vg	(%)Vbg	(%)Vtg	(%)Vbtg	(%)Vgbtg	(%)ort
NB	66.5	69.4	70.2	78.1	78.1	71.7
SVM	80	88.1	91.6	92.2	***92.5***	88.9
C 4.5	51.3	56.8	46	61.1	63.5	55.7
RF	48	51.6	42.5	46	45.7	46.8
(%)ort	61.5	66.5	62.6	68.4	70	***65.8***

Kaynak: Amasyalı, Diri ve Türkoğlu, 2006.

Tablo 16: Özellik Azaltılarak Kullanılan 5 Vektörün Sınıflandırma Başarısı

	(%)Vga	(%)Vbga	(%)Vtga	(%)Vbtga	(%)Vgbtga	(%)ort
NB	75.4	78.4	80.2	85.1	85.6	80,9
SVM	70.3	73.3	83.8	88.1	***88.4***	80,8
C 4.5	58.6	63.7	55.9	67.5	74	63,9
RF	69.5	78.3	69	77.6	82	75,3
(%)ort	68.5	73.4	72.2	79.6	82.5	***75,2***

Kaynak: Amasyalı, Diri ve Türkoğlu, 2006.

2.10.1.Özellik Vektörleri

Yazarlara ait dokümanların öğrenme teknikleri kullanılarak sınıflandırılması için, sınıflandırma yöntemleri kullanılarak bir dokümanın yazarını, türünü belirlemede ve yazarın cinsiyetini tahmin etme çalışmalarında veri setinde yer alan dokümanlardan belirli özelliklerin elde edilmesiyle özellik vektörü yapılabilmektedir. Her doküman, seçtiğimiz eşik değerine göre sayısını bizim belirlediğimiz n adet özelliğe sahiptir ve n boyutlu bir vektör ile ifade edilir. Her bir yazarın farklı özelliği vardır. Bu farklı özellikleri ele alarak bir özellik vektörü oluşturulmaktadır. Vecdi Emre Levent ile Banu Diri'nin birlikte yapmış olduğu çalışmada Türkçe gazete köşe yazarlarının yazar tanıma çalışmasının yapılmasını sağlamışlardır. Çalışmayı yaparken yazarlara ait olan 16 özellik

(cümle ve kelime sayısı, ortalama kelime ve farklı kelime sayısı, nokta, virgül, satır, noktalı virgül, soru işareti, ünlem, isim, fiil, sıfat, zamir, edat ve bağlaç sayısı) belirlenmiştir. Aynı ve farklı kategorilerde yazan yazarlardan ve ayrıca cinsiyete göre üç farklı veri seti kullanılarak yazar tanıma üzerine çalışılmıştır. Çalışmada yapay sinir ağları yöntemi kullanılmıştır. Cinsiyete göre çıkarılmış özellik vektöründe kadınların ve erkeklerin farklı karakteristik özelliklerinin olduğu ortaya çıkarılmıştır.

ÜÇÜNCÜ BÖLÜM
ANALİZ VE YORUM

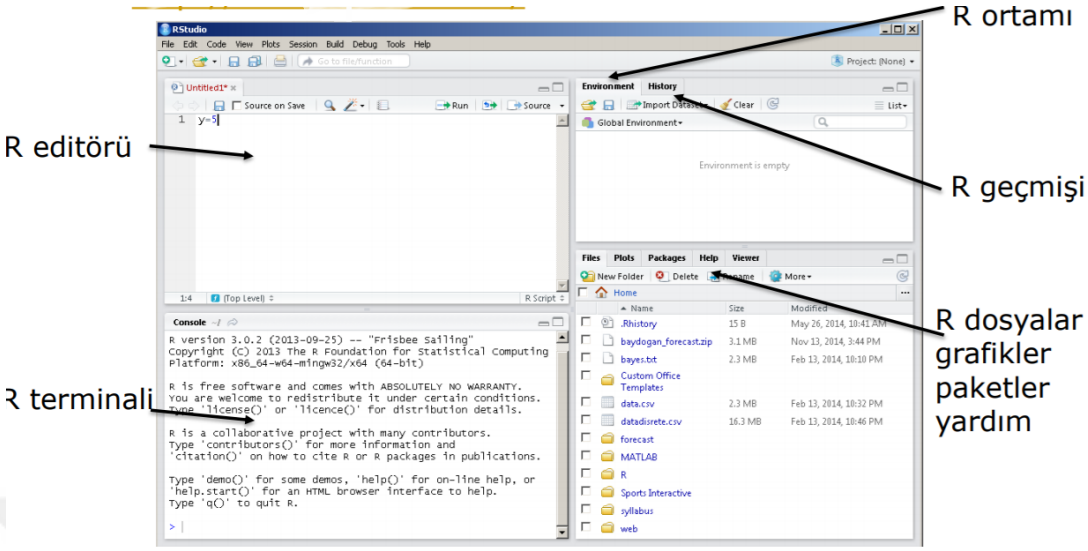
3.1. R PROGRAMLAMA DİLİ

1991 yılında S uyarlaması olarak Ross Ihaka ve Robert Gentleman tarafından Yeni Zelanda da geliştirilen R, güncel olarak 2 milyondan fazla kullanıcı tarafından kullanılmaktadır. Oldukça hızlı bir şekilde büyümeye devam eden dil 2019 itibariyle popüler programlama dilleri arasında (Java, C, Pyhton, C++, Visual Basic .NET, JavaScript, C#, PHP, SQL, Objective-C, MATLAB) 12. sırada yer almaktadır**.** (TIOBE Programlama Topluluk Endeksi'nden gelen verilere dayanmaktadır.)

R (R Project), açık kaynaklı, istatistiksel hesaplama ve grafikler konusunda özelleştirilmiş bir programlama dilidir. Açık kaynaklı olması nedeniyle geliştirilmeye müsait bir programlama dilidir. Doğrusal ve doğrusal olmayan modelleme, klasik istatistik testleri, zaman-serileri analizi, sınıflandırma, kümeleme gibi çok geniş istatistiki ve grafiksel yöntemler için etkin programlama yeteneklerini içerir.

R, istatistik ile ekonometrik hesaplamalar, veri manipülasyonu, programlama ve grafiksel gösterimi bir arada sunan bütünleşmiş yazılım ve programlama dilidir.

Şekil 10: R Programı Başlangıç Ara Yüzü



Kaynak: Baydoğan, Çetin ve Orbay, 2014

3.2. R PROGRAMINDA YAZAR TANIMA UYGULAMASI

R programlama dili açık kaynak kodlu olan bir yazılımdır. R' da hazır paketler kullanılarak analizler yapılabilirken, kullanıcıya bağlı olarak ihtiyaca yönelik yeni paketler de oluşturulabilmektedir. Bu analiz kapsamında kullanılacak olan kütüphaneler 2. Bölümde gösterilmiştir.

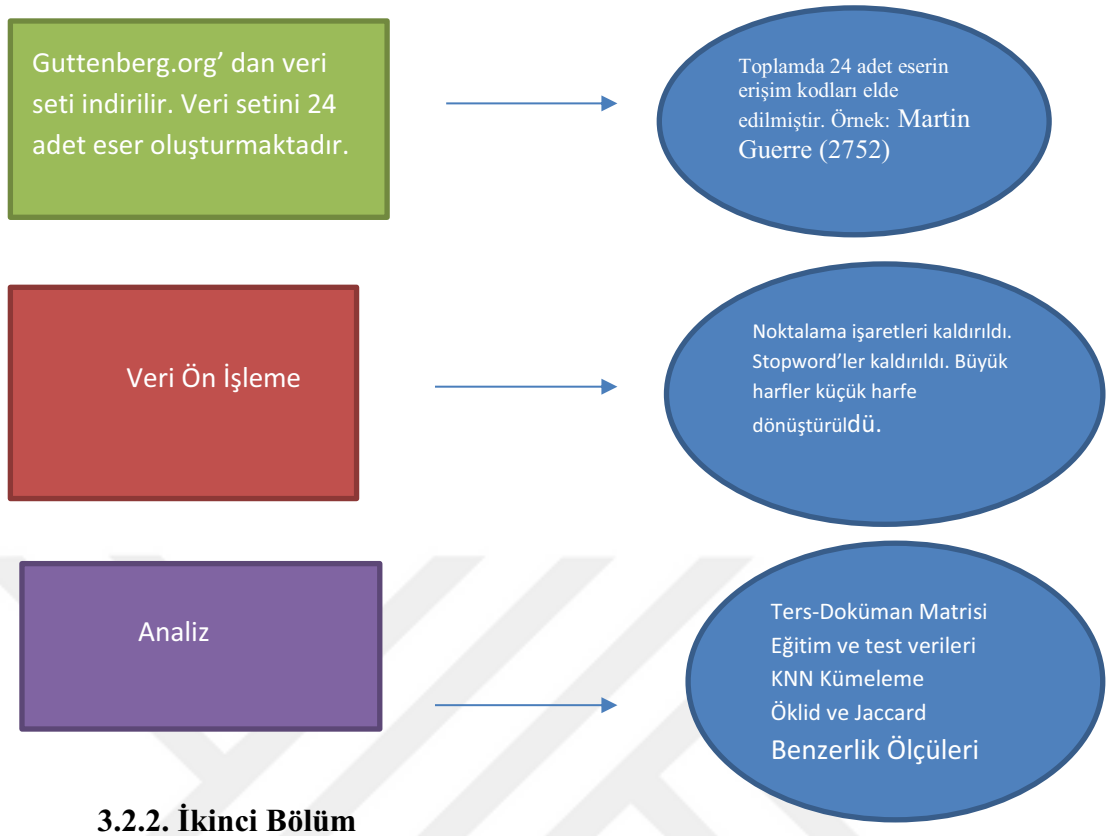
Guttenberg web sitesinden Tolstoy, Jane Austen, Mark Twain ve Alexandre Dumas' a ait 6'şar adet eser alınarak **"term_frequency"** kodu yardımıyla eserlerin en sık tekrar eden 600 kelimesi, **"textrank, udpipe"** kütüphaneleri yardımıyla sıfatları ve zarfları bulunmuştur. En sık tekrar eden kelimeler içerisinden eserler ile ilgili isimler listeden çıkarılmıştır.

İlk olarak Uygulama Süreci bölümünde 4 ayrı veri seti için ayrı ayrı yapılacak olan uygulama süreçleri anlatılmıştır. İkinci ve Üçüncü bölümlerde uygulama süreçleri detaylı olarak ele alınarak analiz sonuçları gösterilmiştir.

3.2.1. Uygulama Süreci

R Studio' da yapılmış olan uygulamanın aşamaları aşağıda detaylı olarak gösterilmiştir.

Şekil 11: R Studio' da yapılmış olan Uygulamanın Aşamaları



3.2.2. İkinci Bölüm

Metin analizi için **tidyverse, tidyr, tm** paketleri **install.package ( )** ile yüklendikten sonra **library ( )** komutu ile çağırılarak kullanılabilir hale getirilir ve **"ggplot"** paketi yardımıyla grafik işlemleri yapılır. Library (gutenberg) komutu yardımıyla gutenberg paketi kütüphaneden çağırılarak yapılan R işlemleri aşağıda gösterilmiştir.

```
library(gutenbergr)

theme_set(theme_minimal())

books <- gutenberg_download( c (2752), meta_fields = "author")

  # 2752 kodlu eser R programına yüklenir.

tidy_books <- books %>%

      unnest_tokens(word, text)

#Eser kelime kelime ayrılır.
```

3.2.3. Metin Ön İşleme

Corpus adlı yapıyı oluşturabilmek için **tm** paketinden **Corpus** ve **VectorSource** paketleri ile işlem yapılmıştır. Daha sonra **Tolower** ile büyük harfler küçük harfe çevrilmiş ve **RemoveNumbers** ile rakamlar kaldırılmıştır. **RemovePunctuation** ile de İngilizce olmayan harfler, karakterler ve tüm noktalama işaretleri metin içerisinden kaldırılmıştır. **Stopwords**, metin içerisinde yer alan tüm durdurma kelimelerin (a, an, the, to vb.) kaldırılmasını sağlarken, **stripWhitespace** ile kelimeler arasında yer alan boşluklar temizlenmiştir. Temizleme işlemleri için **"tm( )"** kütüphanesi içerisinde yer alan **"tm_map( )"** fonksiyonu kullanılmıştır. Metin ön işleme aşamasında dikkat edilmesi gereken en önemli konu işlemlerin sırasıdır. Her çalışmada farklılık gösterebilmektedir. Kesin bir düzeni veya kuralı bulunmamaktadır. Uygulamada kullanılan işlem sırası aşağıda gösterilmektedir.

```
Text_corpus <- Corpus(VectorSource(books))

text_corpus_clean<- tm_map(text_corpus, tolower)

text_corpus_clean<- tm_map(text_corpus_clean, removeNumbers)

text_corpus_clean<- tm_map(text_corpus_clean, removePunctuation)

text_corpus_clean<- tm_map(text_corpus_clean, removeWords, stopwords())

text_corpus_clean<- tm_map(text_corpus_clean, stripWhitespace)
```

3.2.4. Sık Kullanılan Kelimeler

Metin ön işleme aşaması tamamlandıktan sonra üzerinde çalışma yapılan eserle saf veri formatına dönüştürülmüş ve **"TermDocumentMatrix"** fonksiyonu ile **"tdm"** adıyla terim-doküman matrisi oluşturulmuştur. Sık kullanılan kelimelerin bulunulması için öncellikle **"books_m"** matrisi oluşturulmuştur. Ardından matrisindeki toplam satır sayılarının (kelime sayılarının), azalan sıraya göre düzenlenmesiyle **"term_frequency"** listesi oluşturulmuştur.

```
tdm<-TermDocumentMatrix(text_corpus_clean)

books_m<- as.matrix(tdm)

dim(books_m)

 term_frequency<-rowSums(books_m)

term_frequency<-sort(term_frequency, decreasing = TRUE)

term_frequency[1:600]
```

Örnek olarak Alexandre Dumas' ın The Black Tulip adlı eserinin en sık tekrar eden 20 kelimesi gösterilmiştir. The Black Tulip adlı eserde "well" kelimesi 176 kez tekrarlandığı görülürken "prisoners" kelimesi 114 kez tekrarlandığı belirlenmiştir.

Cornelius	**said**	**one**	**will**	**time**	**gryphus**	**well**	**man**	**boxtel**	**Baerle**	**see**	**shall**
613	366	274	267	115	177	176	175	173	171	139	139
Now	**black**	**witt**	**say**	**prisoner**	**yes**	**first**	**two**				
121	120	120	118	114	113	111	107				

3.2.5. Sıfat Ve Zarfların Bulunulması

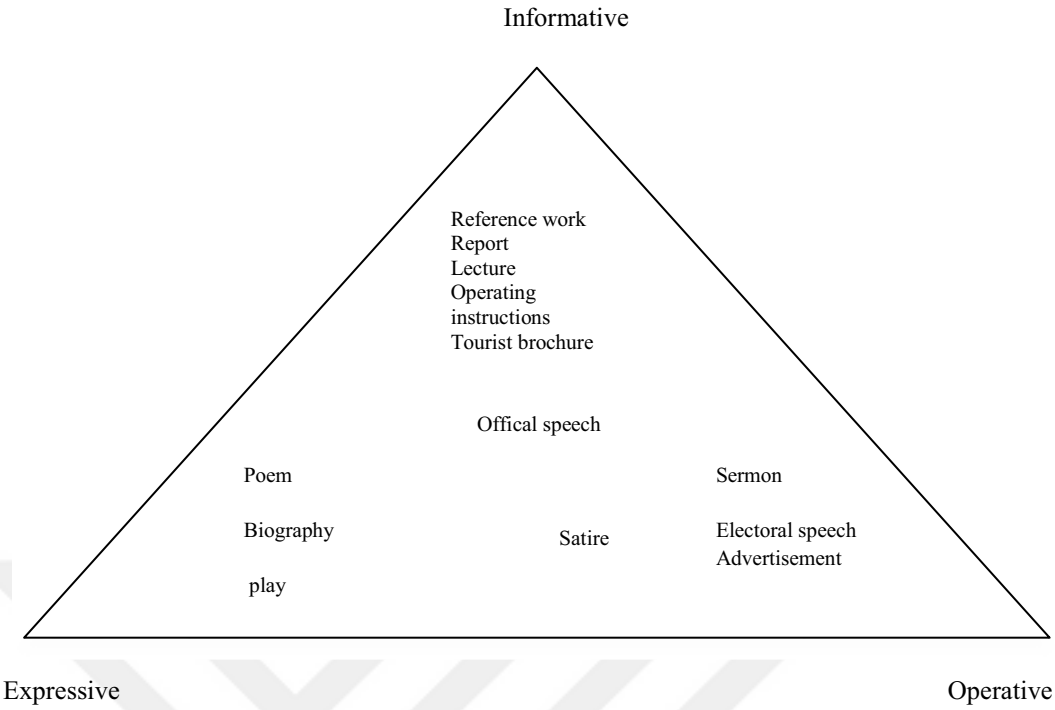
Üzerinde çalışılan eserlerde sıfat ve zarfların bulunması yazarların yazarlık özelliklerinin çıkarılmasında önem taşımaktadır.

R programında yer alan **"textrank"** algoritması, cümlelerin birbirleriyle nasıl ilişkilendirildiğini hesaplayarak metni özetlemeyi sağlar ve metin içerisinden en önemli kelimelerin, anahtar kelimelerin bulunmasını sağlar. **"udpipe"** algoritması ise doğal dil işleme (NLP)' nin bir aracıdır. *Udpipe* sayesinde metin içerisinde *POS Tagging* (Kelimeleri Sıfat, Zarf, Zamir, Bağlaç vb. olarak etiketlemektir.) , *lemmatization* (Kelimenin sözlükteki doğru halini bulmaktır yani çoğul kelimeleri tekil hale getirmek gibi.) *dependency parsing* ("Bağlılık ayrıştırması, bir cümle içerisindeki sözcükler arasındaki ilişkilerin ve ilişki türlerinin belirlenmesidir ve bir cümlenin anlamsal analizinin yapılabilmesi için

şarttır.” (Bilgin ve Amasyalı; 2017: 385)) gibi işlemleri yapmaktadır. Uygulamada *textrank* ve *udpipe* kütüpneleri yüklenmiş olup dil olarak İngilizce sıfat ve zarfların bulunması sağlanmıştır.

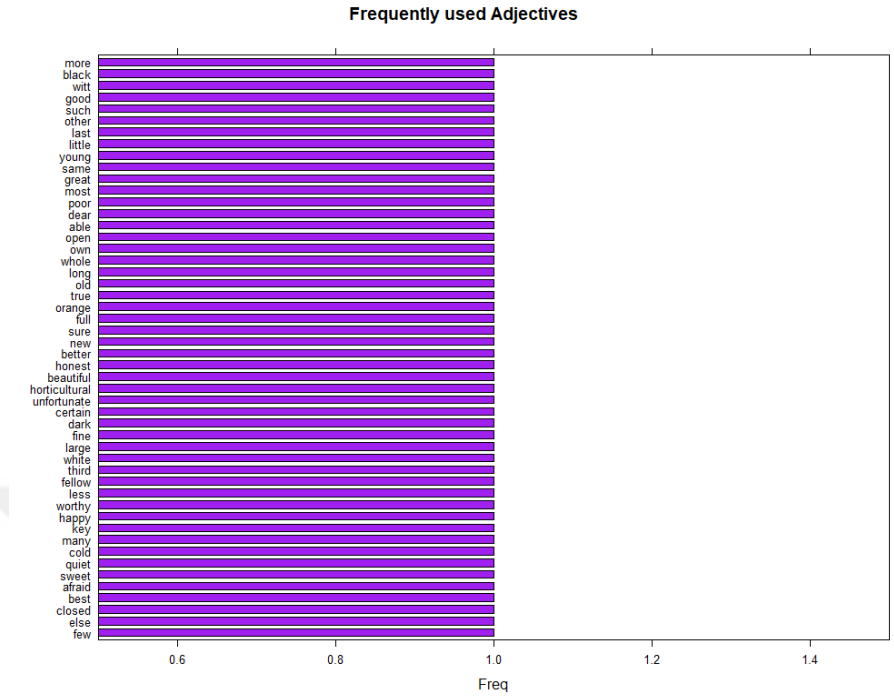
```
library(textrank)

library(udpipe)

ud_model <- udpipe_download_model(language = "english")

ud_model <- udpipe_load_model(ud_model$file_model)

x <- udpipe_annotate(ud_model, x=words$word )

x <- as.data.frame(x)

library(lattice)

adj <- subset(x, upos %in% c("ADJ"))

#Tüm sıfatlar bulunur.

adj <- txt_freq(adj$token)

adj$key <- factor(adj$key, levels = rev(adj$key))

barchart(key ~ freq, data = head(adj, 50), col = "purple",

main = "Frequently used Adjectives", xlab = "Freq")
```

Şekil 12: Reiss's text types an text varieties (Chesterman 1989:105)



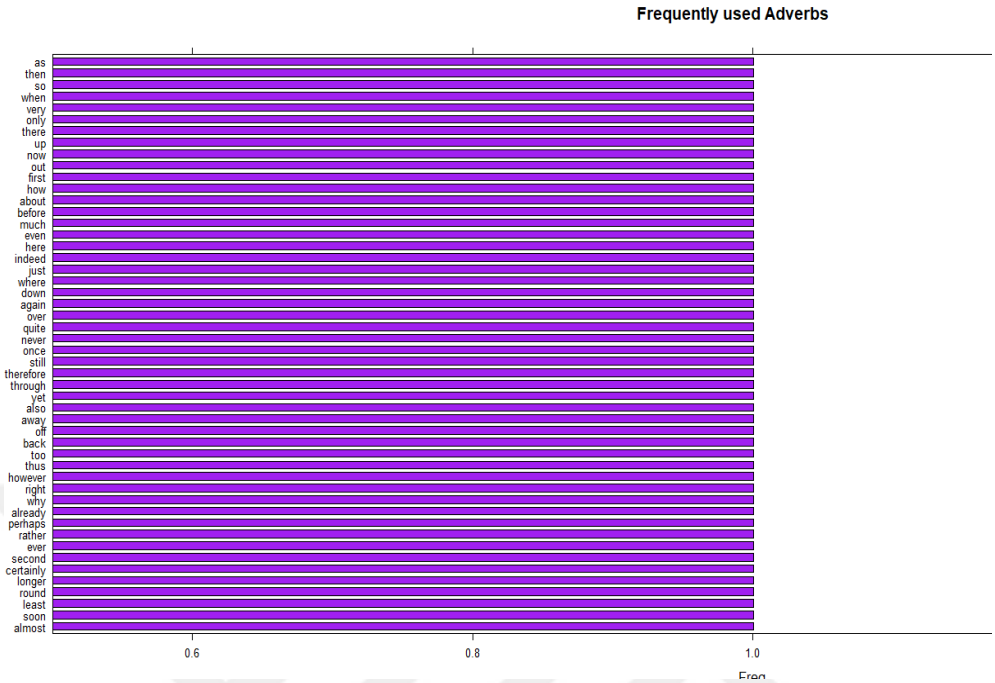
Yukarıda yer alan üçgende referans çalışmalar, raporlar ve turistik broşürler bilgi veren eserler olarak tanımlanırken oyun, şiir, biyografi gibi eserler okuyucu etkileyen eserler olarak tanımlanmaktadır. Bu türde yazan yazarların yazarlık özellikleri çıkarılırken dikkat edilmesi gereken nokta roman, şiir, biyografi gibi eserlerde yazarın kullanmış olduğu sıfat, zarf, zamir, noktalama işaretleri gibi özelliklerin tespit edilmesinin gerekliliğidir. Çünkü bir yazarı diğer yazardan ayıran özellikler içerisinde kullanmış olduğu sıfatların, zarfların çeşidi ve kullanım sıklığı yer alabilmektedir. Çalışmada yer alan eserler roman türünde yazıldığı için her yazarın eserlerinde en sık kullandığı sıfat ve zarflar R programı aracılığı ile bulunmuştur. Aşağıda yer alan Tablo 17' de Alexandre Dumas' ın The Black Tulip adlı romanında en sık kullanmış olduğu ilk 50 sıfat yer almaktadır.

Tablo 17: Alexandre Dumas - The Black Tulip En sık kullanılan ilk 50 Sıfat




```
adv <- subset(x, upos %in% c("ADV"))

#Tüm zarflar bulunur.

adv <- txt_freq(adv$token)

adv$key <- factor(adv$key, levels = rev(adv$key))

barchart(key ~ freq, data = head(adv, 50), col = "purple",

        main = "Frequently used Adverbs", xlab = "Freq")
```

Şekil 13: Alexandre Dumas - The Black Tulip En sık kullanılan ilk 50 Zarf



Metin ön işleme aşamaları tüm eserlere uygulandıktan sonra çıkan sonuçlar .csv formatında kaydedilmiştir. Böylece yazarlara ait kişilik özellikleri elde edilmiş olup analiz aşamasında test edilecektir.

3.3. ANALİZ

Analiz bölümünde Ters-doküman matrisi, eğitim ve test verileri, KNN Kümeleme, Öklid ve Jaccard benzerlik ölçüleri ele alınmıştır.

3.3.1. Terim-Doküman Matrisi

Doküman-terim matrisi, bir belge topluluğunda oluşan terimlerin sıklığını tanımlayan matematiksel bir matristir. Bir terim-doküman matrisinde, satırlar koleksiyondaki terimlere karşılık gelir ve sütunlar dokümanlara karşılık gelmektedir. Doküman-terim matrisinde ise durum tam tersidir. Bu matrislerin oluşturulma amacı terimleri ve aralarındaki ilişkileri incelemek ve görselleştirme işlemini kolaylaştırmaktır.

```
[[1]]
[[1]]$name
[1] "dumas"

[[1]]$tdm
<<TermDocumentMatrix (terms: 497, documents: 6)>>
Non-/sparse entries: 1369/1613
Sparsity           : 54%
Maximal term length: 14
Weighting          : term frequency (tf)


[[2]]
[[2]]$name
[1] "jane"

[[2]]$tdm
<<TermDocumentMatrix (terms: 590, documents: 6)>>
Non-/sparse entries: 2319/1221
Sparsity           : 34%
Maximal term length: 16
Weighting          : term frequency (tf)


[[3]]
[[3]]$name
[1] "mark"

[[3]]$tdm
<<TermDocumentMatrix (terms: 660, documents: 6)>>
Non-/sparse entries: 2005/1955
Sparsity           : 49%
Maximal term length: 11
Weighting          : term frequency (tf)


[[4]]
[[4]]$name
[1] "tolstoy"

[[4]]$tdm
<<TermDocumentMatrix (terms: 530, documents: 6)>>
Non-/sparse entries: 1333/1847
Sparsity           : 58%
Maximal term length: 13
Weighting          : term frequency (tf)
```

Yukarıdaki sonuçlardan da görüleceği üzere Tolstoy veri kümesi 530 satır ve 6 dokümandan(sütun) oluşmaktadır.

3.3.2. Eğitim Ve Test Verisi Oluşturma

"Makine öğrenmesi tüm üç farklı yöntemle çalışabilmektedir. Bunlar; denetimli (gözetimli) öğrenme, denetimsiz (gözetimsiz) öğrenme ve yarı denetimli öğrenmedir. Denetimli öğrenme; sisteme eğitim veri seti ve test veri setinin yüklenmesi, veri setinde her bir veri için gerekli etiketlenmenin yapılması ve bu sayede girdi veri seti ile çıktı veri seti arasında ilişki kurulması mantığına dayanır. Temel amaç sonuçları bilinen veri setinden yapılan sınıflandırmadan hareketle sonuçları bilinmeyen veri setine ilişkin etkili tahminler yapabilmektir (Aydın ve Özkul, 2015:38). "

"Denetimsiz öğrenme de ise kullanıcının sisteme herhangi bir müdahalesi yoktur. Sadece girdi verileri sisteme verilir ancak herhangi bir işaretleme yapılmaz. Sistem otomatik olarak keşifler yapar, ilişki ağını ortaya koymaya çalışır (Alpaydın, 2010:11)."

Çalışmada yarı denetimli öğrenme yöntemi olarak ele alınmış, veri setinin %70'lik kısmı eğitim %30'luk kısmı test verisi olarak kullanılmıştır.

```
novTDM <- lapply(tdm, bindnovelTDM)

str(novTDM)

tdm.stack <- do.call(rbind.fill , novTDM)

tdm.stack[is.na(tdm.stack)] <- 0

head(tdm.stack)

train.idx <- sample(nrow(tdm.stack), ceiling(nrow(tdm.stack) * 0.7))

test.idx <- (1:nrow(tdm.stack))[-train.idx]
```

3.3.3. K-EN Yakın Komşu (KNN) Algoritması

"K-En yakın komşu algoritması makine öğrenim algoritmaları içerisinde en çok bilinen ve kullanılan sınıflandırma algoritmalardan birisidir. Seçilen bir özelliğin kendine en yakın olan özellikle arasındaki yakınlığı kullanarak sınıflandırma yapılır. Çalışma şekline baktığımızda, tanımlanan verilere göre yeni bir tanımlanması gereken nesne geldiğinde öncelikle K değerine bakılır. Burada eşitlik olmaması için genellikle K sayısı tek sayı olarak seçilir. Yeni gelen veri ile diğer veriler arasındaki mesafeler hesaplanırken Kosinüs, Öklid ya da Manhattan uzaklığı gibi yöntemler kullanılır." (Kılınç, Borandağ, Yücalar, Tunalı, Şimşek ve Özçift; 2016: 90)

Aşağıdaki Tablo 18 Terim-Doküman Matrisine bakıldığında "Answer" kelimesi 1 nolu eserde geçiyor ise 1 geçmiyorsa 0 değerini almaktadır. (Bknz. Ek.1.)

Tablo 18: Terim-Doküman Matrisi

Filter

	answer	back	became	came	can	continued	cried	ever	exclaimed	father	fell	give	head	idea	judge	less
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
2	1	0	1	1	0	0	1	1	0	1	1	0	0	1	0	
3	1	1	0	1	1	1	0	1	1	0	1	1	1	1	1	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	1	0	1	0	0	0	0	1	1	0	1	0	0	0	0	
6	1	0	0	0	0	0	0	1	0	1	1	0	0	1	0	
7	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	
8	1	0	1	0	0	1	0	0	0	0	0	0	1	1	0	
9	1	1	1	0	0	1	1	0	0	0	0	0	1	1	0	
10	1	0	1	0	0	1	1	0	0	0	0	0	1	1	0	
11	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	
12	1	1	0	0	0	1	0	0	0	0	0	0	1	1	0	
13	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
14	1	0	0	0	0	0	1	0	0	0	1	0	0	1	1	
15	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	
16	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	
17	1	1	1	1	0	1	0	1	0	0	1	1	1	1	1	
18	1	1	1	1	0	1	1	0	0	0	1	1	1	0	0	

Sınıflandırma algoritmalarının başarımının değerlendirilmesi için Hata Matrisi (Confusion Matrix) kullanılmaktadır. Tablo-19'da yer verilen Hata Matrisi'nde, DP "Doğru Pozitif", YP "Yanlış Pozitif", YN "Yanlış Negatif", DN ise "Doğru Negatif" değerini ifade etmektedir.

Tablo 19: Hata Matrisi

		Tahmin Edilen Sınıf	
		S1(+)	S2(+)
Gerçek Sınıf	S1(+)	DP	YN
	S2(+)	YP	DN

Doğruluk oranı (ACC), sınıflandırıcının sınıflama başarısını ölçmek için geniş çapta kullanılan bir ölçüdür. Yapılan deneysel çalışmada KNN algoritması kullanılmış olup doğruluk oranı(ACC) %85 olarak çıkmıştır.

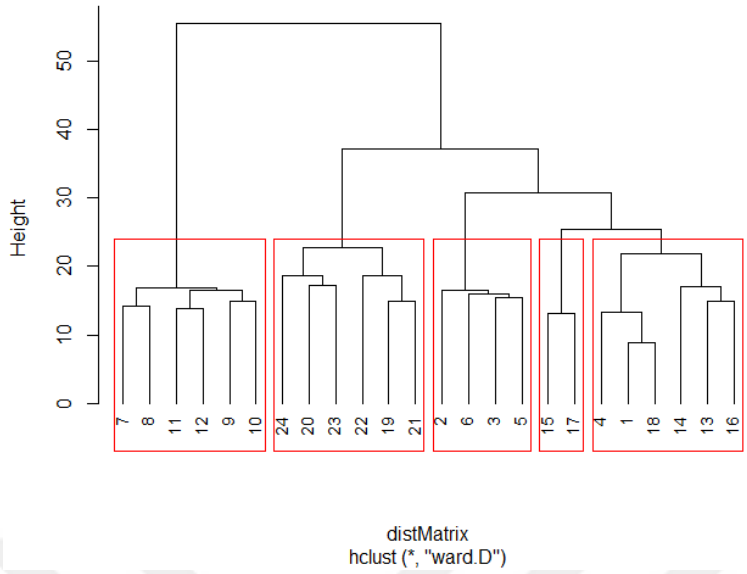
```
tdm.nov <- tdm.stack[, "targetnovel"]

tdm.stack.nl <- tdm.stack[, !colnames(tdm.stack) %in% "targetnovel"]

knn.pred <- knn(tdm.stack.nl[train.idx, ], tdm.stack.nl[test.idx, ],
tdm.nov[train.idx])conf.mat <- table("Predictions" = knn.pred , Actual =
tdm.nov[test.idx])

accuracy <- sum(diag(conf.mat)) / length(test.idx) * 100

accuracy

0,85
```

3.3.4. Öklid Ve Jaccard Benzerlik Ölçüleri

Bu deneysel çalışmada Öklid ve Jaccard Benzerlik ölçüleri ayrı ayrı hesaplanmış olup sonuçlar aşağıda yer almaktadır. Küme sayısı (K) = 5 alınarak Öklid uzaklık ölçüsü yardımıyla WARD kümeleme yöntemi kullanılmıştır.

Jaccard Benzerlik Ölçüsü, kümeler arasındaki benzerliği istatistiksel olarak hesaplamaktadır. Jaccard Benzerliği, iki dokümandaki kelimelerin kesişiminin iki dokümandaki kelime sayısının birleşimine bölünmesi ile elde edilmektedir.

Şekil 14: Cluster Dendrogram



Öklid uzaklık ölçüsü kullanılarak eserler arasındaki benzerlik hesaplanmıştır. Öklid ölçüsü yardımıyla K=5 alınarak birbirine en yakın olan eserler Ward tekniği ile kümelenmiştir. Burada ilk kümede 7. Eser Jane Austen'ın Emma adlı eseridir. 8, 9,10, 11, 12 nolu eserler de Jane Austen'a ait olduğu için ilk küme doğru sınıflandırılmıştır. 5. Kümede ise Alexandre Dumas'ın 1. Ve 4. Eseri ile Tolstoy'un eserleri arasında benzerlik olduğu gözlemlenmiştir. Öklid uzaklığı ile yapılan Ward kümeleme algoritmasının sınıflandırma başarısı %65 olarak bulunuluştur.

```
cluster  <- as.matrix(tdm.stack.nl)
distMatrix <- dist(m, method="euclidean")
groups <- hclust(distMatrix,method="ward.D")
plot(groups, cex=0.9, hang=-1)
rect.hclust(groups, k=5)
```

Jaccard Benzerlik ölçüsü sonuçları;

Jaccard benzerlik ölçüsü, iki kümenin birbiri ile ne kadar benzediğini gösteren ve 0 ile 1 arası bir değer alan benzerlik ölçüsüdür.

```
library(philentropy)

Distance(tdm.stack.nl, method= "jaccard")
```

Aşağıda yer alan Jaccard Benzerlik ölçüsü tablosunda eserler arasındaki benzerlikler yer almaktadır. V1(Martin Guerre) ile V24 (War And Peace) adlı eser %96 oranında birbirine benzemektedir. Öklid ölçüsü kullanılarak yapılmış olan Ward kümeleme sonucunda da Martin Guerre ve War and peace aynı küme içerisinde yer almaktadır. Benzerlik oranı %85 olarak alınmış olup %85'in altındakiler kırmızı ile belirtilmiştir. Jaccard benzerlik matrisi sonuçları incelendiğinde yapılan deneysel çalışma için Öklid uzaklık ölçüsünün, Jaccard uzaklık ölçüsüne göre daha anlamlı sonuçlar verdiği görülmüştür.

Şekil 15: Jaccard Benzerlik Matrisi

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
v1	0.0000000	0.9464883	0.9146341	0.9805195	0.9691877	0.9606557	0.9689119	0.9635417	0.9537037	0.9590909	0.9646465
v2	0.9464883	0.0000000	0.6417234	0.8405797	0.5623529	0.6506024	0.8195212	0.7973734	0.7522442	0.7614841	0.7996324
v3	0.9146341	0.6417234	0.0000000	0.8394737	0.5348315	0.6120092	0.7950530	0.7777778	0.7110333	0.7359455	0.7738516
v4	0.9805195	0.8405797	0.8394737	0.0000000	0.7984085	0.7927928	0.9407895	0.9266667	0.9260000	0.9289941	0.9331897
v5	0.9691877	0.5623529	0.5348315	0.7984085	0.0000000	0.5644028	0.8296796	0.8156997	0.7697368	0.7819063	0.8154362
v6	0.9606557	0.6506024	0.6120092	0.7927928	0.5644028	0.0000000	0.8456014	0.8223443	0.7962003	0.7920962	0.8365897
v7	0.9689119	0.8195212	0.7950530	0.9407895	0.8296796	0.8456014	0.0000000	0.4329004	0.5180952	0.4413519	0.5010183
v8	0.9635417	0.7973734	0.7777778	0.9266667	0.8156997	0.8223443	0.4329004	0.0000000	0.4893204	0.4475248	0.4918033
v9	0.9537037	0.7522442	0.7110333	0.9260000	0.7697368	0.7962003	0.5180952	0.4893204	0.0000000	0.4188679	0.4448819
v10	0.9590909	0.7614841	0.7359455	0.9289941	0.7819063	0.7920962	0.4413519	0.4475248	0.4188679	0.0000000	0.4924242
v11	0.9646465	0.7996324	0.7738516	0.9331897	0.8154362	0.8365897	0.5010183	0.4918033	0.4448819	0.4924242	0.0000000
v12	0.9577114	0.8007246	0.7535461	0.9230769	0.7962963	0.8206039	0.5010060	0.4607438	0.4394531	0.4578544	0.4029536
v13	0.9813084	0.8508065	0.7996071	0.8954424	0.8087954	0.8254620	0.8360071	0.8401421	0.8157454	0.8076923	0.8494810
v14	0.9595376	0.8622642	0.7962617	0.9302885	0.8409894	0.8432122	0.8566667	0.8547579	0.8235294	0.8177496	0.8462810
v15	0.9685714	0.8384615	0.7737643	0.9148418	0.7989031	0.8047337	0.8134715	0.8114187	0.7951220	0.7733990	0.8330551
v16	0.9519520	0.8369352	0.7636719	0.9095477	0.8165138	0.8375734	0.8461538	0.8382100	0.7970297	0.7970540	0.8276451
v17	0.9458824	0.8444816	0.7422680	0.9297189	0.8227848	0.8352941	0.8464978	0.8272727	0.7930029	0.7735294	0.8389956
v18	0.9232955	0.8308271	0.7371429	0.9209302	0.7989324	0.8204159	0.8366337	0.8327815	0.7933227	0.7818471	0.8398058
v19	0.9701493	0.8778055	0.8607889	0.9458484	0.8489703	0.8756219	0.8802521	0.8964803	0.8667954	0.8704762	0.8945233
v20	0.9404762	0.7649485	0.7070707	0.9029851	0.7315175	0.7734694	0.7992958	0.7864769	0.7332186	0.7228916	0.7673179
v21	0.9543974	0.8395062	0.8297872	0.9399478	0.8380414	0.8611670	0.8425760	0.8446429	0.8238255	0.8157454	0.8168761
v22	0.9701493	0.8310811	0.8110403	0.9027356	0.7881356	0.8022989	0.8219178	0.8310680	0.8137432	0.7963636	0.8544776
v23	0.9589041	0.8556485	0.8447937	0.9342466	0.8505747	0.8818737	0.8566243	0.8729875	0.8484848	0.8324873	0.8494624
v24	0.9634146	0.9587302	0.9517045	0.9375000	0.9475138	0.9687500	0.9568528	0.9541985	0.9550562	0.9463087	0.9477612

	v12	v13	v14	v15	v16	v17	v18	v19	v20	v21	v22
v1	0.9577114	0.9813084	0.9595376	0.9685714	0.9519520	0.9458824	0.9232955	0.9701493	0.9404762	0.9543974	0.9701493
v2	0.8007246	0.8508065	0.8622642	0.8384615	0.8369352	0.8444816	0.8308271	0.8778055	0.7649485	0.8395062	0.8310811
v3	0.7535461	0.7996071	0.7962617	0.7737643	0.7636719	0.7422680	0.7371429	0.8607889	0.7070707	0.8297872	0.8110403
v4	0.9230769	0.8954424	0.9302885	0.9148418	0.9095477	0.9297189	0.9209302	0.9458484	0.9029851	0.9399478	0.9027356
v5	0.7962963	0.8087954	0.8409894	0.7989031	0.8165138	0.8227848	0.7989324	0.8489703	0.7315175	0.8380414	0.7881356
v6	0.8206039	0.8254620	0.8432122	0.8047337	0.8375734	0.8352941	0.8204159	0.8756219	0.7734694	0.8611670	0.8022989
v7	0.5010060	0.8360071	0.8566667	0.8134715	0.8461538	0.8464978	0.8366337	0.8802521	0.7992958	0.8425760	0.8219178
v8	0.4607438	0.8401421	0.8547579	0.8114187	0.8382100	0.8272727	0.8327815	0.8964803	0.7864769	0.8446429	0.8310680
v9	0.4394531	0.8157454	0.8235294	0.7951220	0.7970297	0.7930029	0.7933227	0.8667954	0.7332186	0.8238255	0.8137432
v10	0.4578544	0.8076923	0.8177496	0.7733990	0.7970540	0.7735294	0.7818471	0.8704762	0.7228916	0.8157454	0.7963636
v11	0.4029536	0.8494810	0.8462810	0.8330551	0.8276451	0.8389956	0.8398058	0.8945233	0.7673179	0.8168761	0.8544776
v12	0.0000000	0.8339100	0.8447712	0.8120805	0.8183362	0.8116592	0.8078818	0.8964143	0.7513321	0.8382609	0.8200758
v13	0.8339100	0.0000000	0.7448980	0.5333333	0.6954644	0.7029520	0.7519685	0.8648649	0.7828685	0.8548708	0.8120805
v14	0.8447712	0.7448980	0.0000000	0.6728016	0.7361111	0.5736434	0.7098646	0.9000000	0.8161765	0.8947368	0.8606061
v15	0.8120805	0.5333333	0.6728016	0.0000000	0.6540084	0.7001764	0.7054264	0.8623853	0.7853107	0.8576779	0.8179916
v16	0.8183362	0.6954644	0.7361111	0.6540084	0.0000000	0.6697248	0.6958250	0.9123596	0.7873563	0.8500000	0.8597938
v17	0.8116592	0.7029520	0.5736434	0.7001764	0.6697248	0.0000000	0.5700758	0.9020716	0.7800000	0.8721683	0.8521127
v18	0.8078818	0.7519685	0.7098646	0.7054264	0.6958250	0.5700758	0.0000000	0.8969957	0.7584270	0.8370370	0.8269618
v19	0.8964143	0.8648649	0.9000000	0.8623853	0.9123596	0.9020716	0.8969957	0.0000000	0.7279793	0.8823529	0.7034700
v20	0.7513321	0.7828685	0.8161765	0.7853107	0.7873563	0.7800000	0.7584270	0.7279793	0.0000000	0.6876356	0.6374696
v21	0.8382609	0.8548708	0.8947368	0.8576779	0.8500000	0.8721683	0.8370370	0.8823529	0.6876356	0.0000000	0.8202247
v22	0.8200758	0.8120805	0.8606061	0.8179916	0.8597938	0.8521127	0.8269618	0.7034700	0.6374696	0.8202247	0.0000000
v23	0.8457447	0.8378378	0.8787879	0.8574181	0.8674464	0.8590604	0.8536585	0.8685567	0.7462687	0.4807692	0.8103044
v24	0.9514563	0.9602446	0.9610028	0.9553073	0.9715909	0.9709821	0.9656992	0.8944724	0.9115044	0.9529781	0.9053030

	v23	v24
v1	0.9589041	0.9634146
v2	0.8556485	0.9587302
v3	0.8447937	0.9517045
v4	0.9342466	0.9375000
v5	0.8505747	0.9475138
v6	0.8818737	0.9687500
v7	0.8566243	0.9568528
v8	0.8729875	0.9541985
v9	0.8484848	0.9550562
v10	0.8324873	0.9463087
v11	0.8494624	0.9477612
v12	0.8457447	0.9514563
v13	0.8378378	0.9602446
v14	0.8787879	0.9610028
v15	0.8574181	0.9553073
v16	0.8674464	0.9715909
v17	0.8590604	0.9709821
v18	0.8536585	0.9656992
v19	0.8685567	0.8944724
v20	0.7462687	0.9115044
v21	0.4807692	0.9529781
v22	0.8103044	0.9053030
v23	0.0000000	0.9503311
v24	0.9503311	0.0000000

3.3.5. Duygu Analizi (SENTİMENT ANALYSİS)

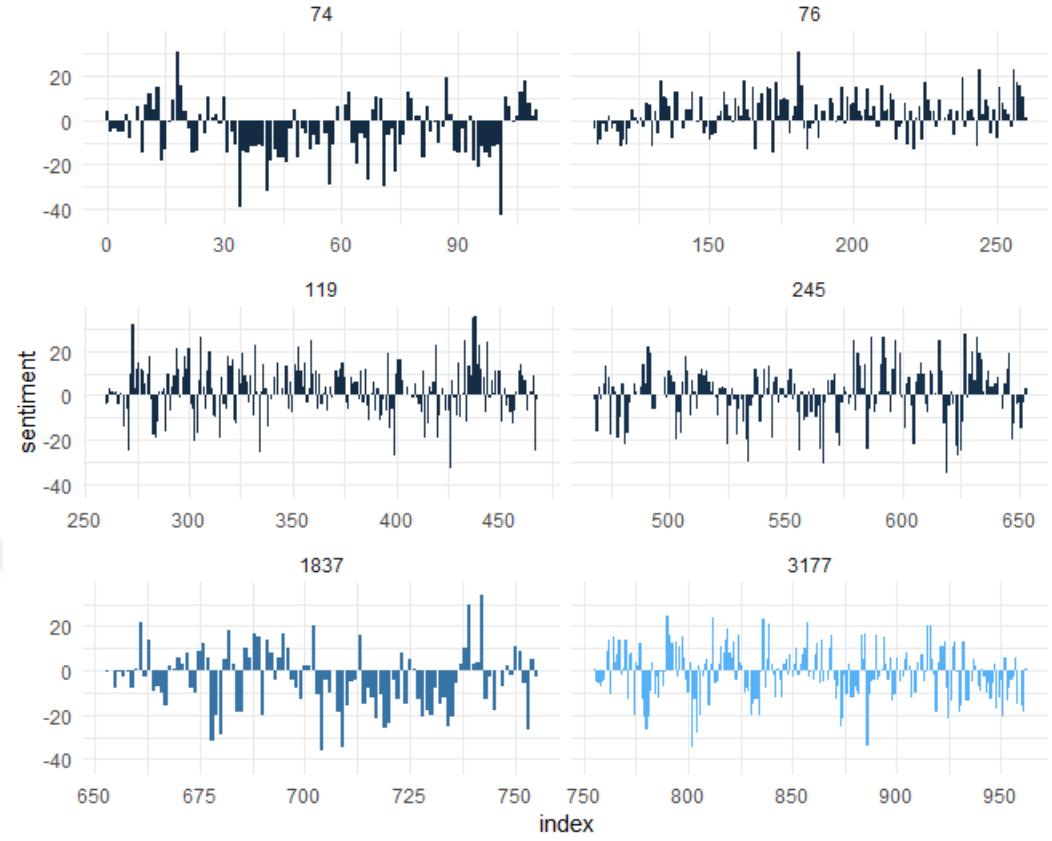
Bu bölümde her bir yazara ait eserlerden yola çıkılarak kullanılan kelimeler pozitif, negatif, nötr olarak sınıflandırılmıştır. Duygu analizinde amaç metinlerde kullanılan dil hakkında bilgi sahibi olmaktadır.

Öncelikle Mark Twain' e ait olan 6 eseri (A Tramp Abroad, Adventures of Huckleberry Finn, Life on the Mississippi, Roughing it, The Adventures of Tom Sawyer, The Prince and The Pauper) ele alarak kelimeleri çıkartıldı. Duygu analizi için gerekli olan pozitif ve negatif kelimeler için 3 sözlük (BİNG, AFINN, NRC) bulunmaktadır. Araştırma kapsamında "BİNG" sözlüğünden yararlanılacaktır. Bu sözlüklerin hepsi unigram yani tek sözcüklere dayanmaktadır. Sözcükleri tek tek ele almaktadır. "**Afinn** sözlüğü olumlu / olumsuz şeklinde her bir kelimenin -5 ile +5 arasında sayısal skoru bulunmaktadır. **Bing** sözlüğünde olumlu / olumsuz şeklinde iki kategoride bulunur. **Nrc** sözlüğünde ise pozitif, negatif, öfke, beklenti, tiksinti, korku, sevinç, üzüntü, şaşkınlık ve güven (positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise ve trust) gibi kategoriler bulunur." (*Julia Silge; David Robinson; 2019; Text Mining with R)*

Duygu analizi çalışmasında R programında yer alan ***"tidyr", "ggplot2", "dplyr"*** *ve* ***"stringr"*** paketleri kullanılmıştır.

```
mark_twain_sentiment <- tidy_books %>%

inner_join(get_sentiments("bing")) %>%

count(gutenberg_id, index = linenumber %/% 80, sentiment)

spread(sentiment, n, fill = 0) %>%

mutate(sentiment = positive - negative)
```

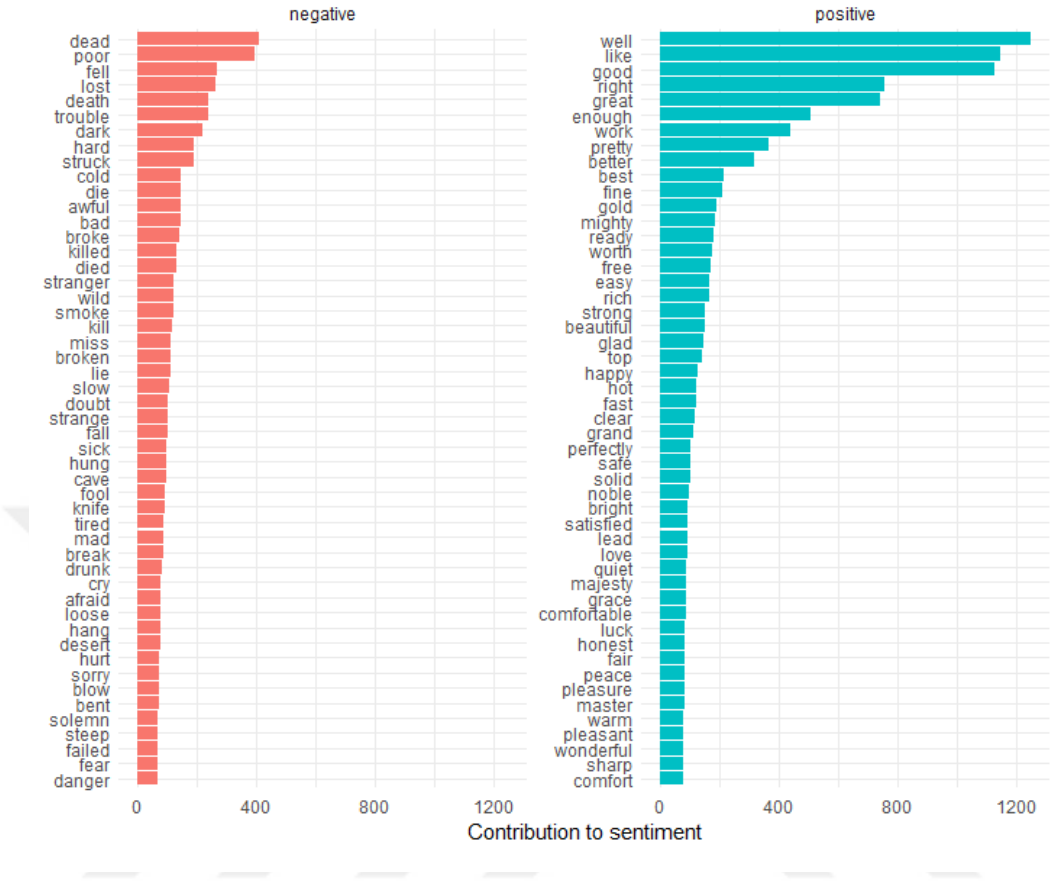
Şekil 16: Mark Twain Eserleri İçin Duygu Analizi Frekans Tablosu



Yukarıda 6 esere ait olan negatif ve pozitif sözcüklerin frekansları yer almaktadır. **74 = The Adventures of Tom Sawyer** adlı eserde daha çok negatif anlam taşıyan kelimeler kullanılmakta iken **76 = Adventures of Huckleberry Finn** adlı eserde pozitif anlam taşıyan kelimeler çoğunluktadır. (Ek.2.)

Yazarın eserlerinde en yaygın kullandığı pozitif ve negatif 50 kelimeyi Bing sözlüğü ile ele alalım. Aşağıdaki tabloda görüleceği üzere ölüm, korku, kaybetmek, kırılmak gibi kelimeler olumsuz olarak tespit edilmiş iken iyi, güzel, zengin, güçlü gibi kelimeler de pozitif olarak belirlenmiştir. Yapılan çalışma sonucunda Mark Twain ile ilgili eserlerinde her ne kadar olumlu bir dil kullanmış olsa da zengin-fakir, özgürlük, ölüm gibi ifadelerle de toplumsal sorunları eleştirel bir bakış açısıyla ele aldığını göstermektedir. Örnek olarak The Adventures Of Tom Sawyer her ne kadar bir macera kitabı olarak gözükse de aslında köleliği, insan doğasının ikiyüzlülüğünü de ele almaktadır.

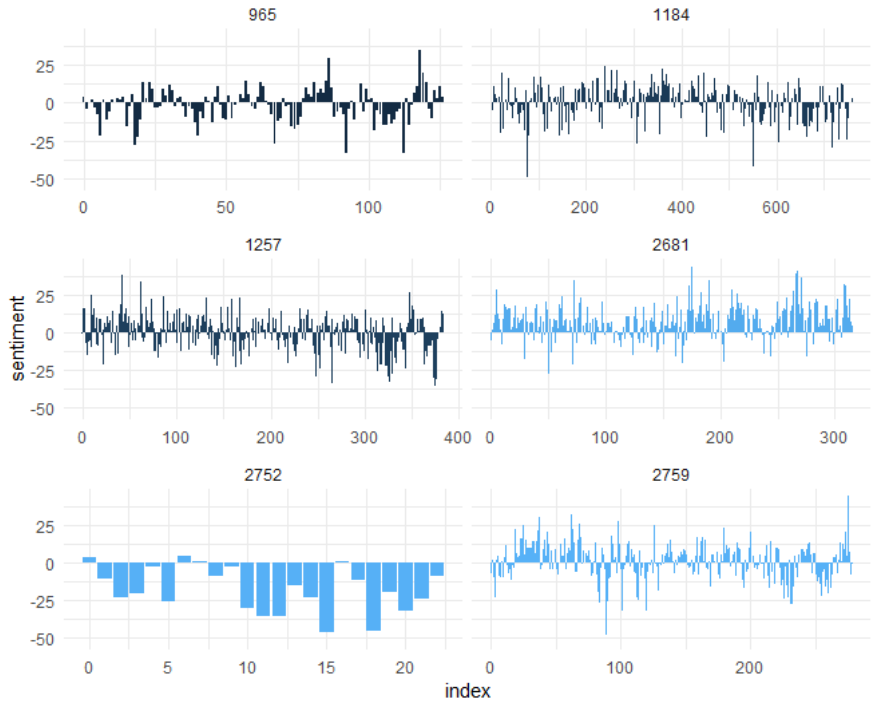
Şekil 17: Mark Twain Eserleri İçin Positive Ve Negative Kelimeler



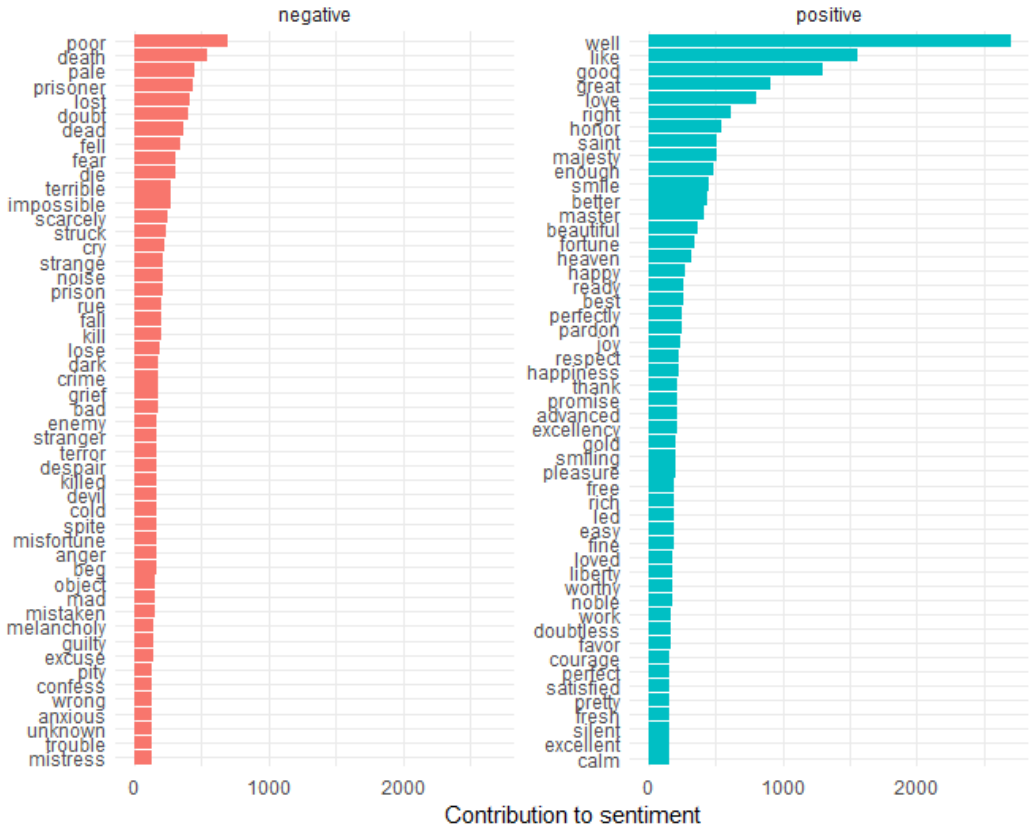
Alexandre Dumas için yapılan duygu analizi çalışma sonuçları aşağıda yer almaktadır. Çıkan grafikler incelendiğinde Ten Years Later(2681) adlı eser daha çok pozitif kelimelerden oluşurken Martin Guerre (2752) adlı eser ise negatif kelimelerden oluşmaktadır. 6 eseri toplu olarak incelersek Alexandre Dumas eserlerini yazarken pozitif bir dil kullandığını ve okurken okuyucuda olumlu etkiler bırakmayı hedeflediği sonucunu çıkarabilmekteyiz. Martin Guerre polisiye tarzda yazılan bir eserdir. Eserde çok fazla ölüm, idam gibi olumsuz kelimeler yer almaktadır. Yapılan duygu analizi çalışması da Martin Guerre isimli eserin negatif dile sahip olduğunu gösterir niteliktedir.

Alexandre Dumas, macera dilindeki tarihi romanları ile ünlü olan bir yazardır. Tablo.18 ‘ e bakıldığında yazar death, prisoner, lost, crime gibi olumsuz kelimeleri yaygın olarak kullanırken love, happy, excellent ve joy gibi olumlu kelimeleri de kullanmaktadır.

Şekil 18: Alexandre Dumas Eserleri İçin Duygu Analizi Frekans Tablosu

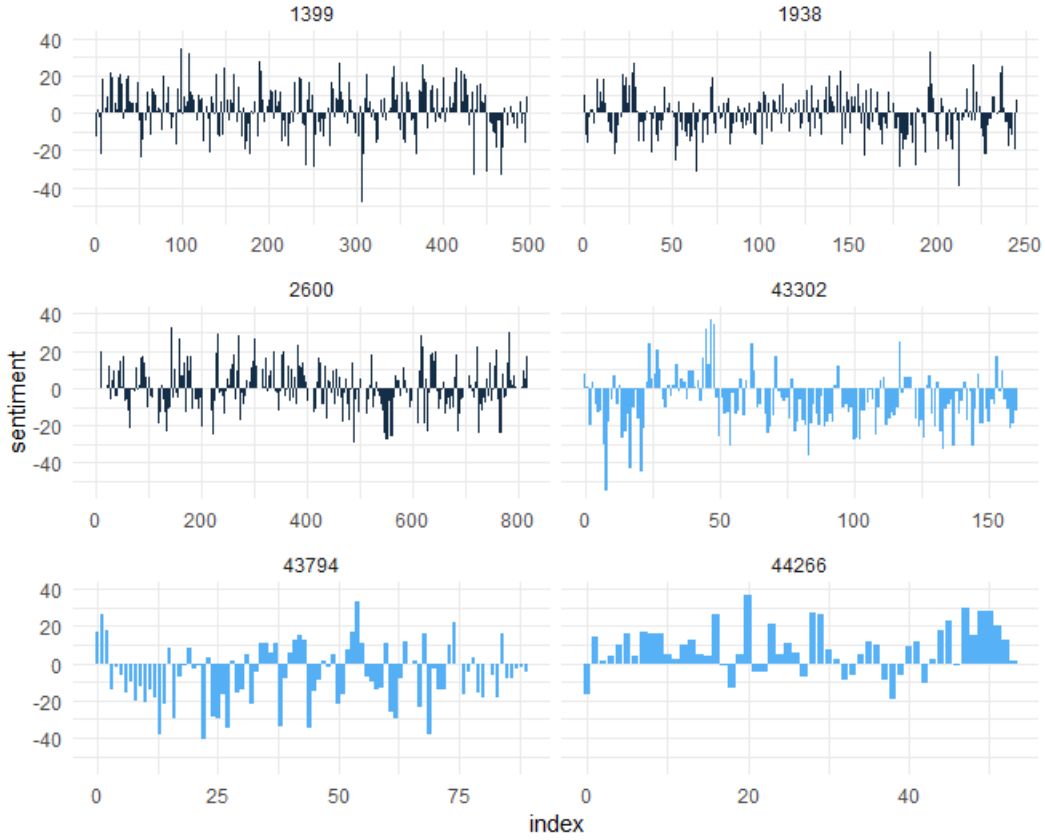


Şekil 19: Alexandre Dumas Eserleri İçin Positive Ve Negative Kelimeler



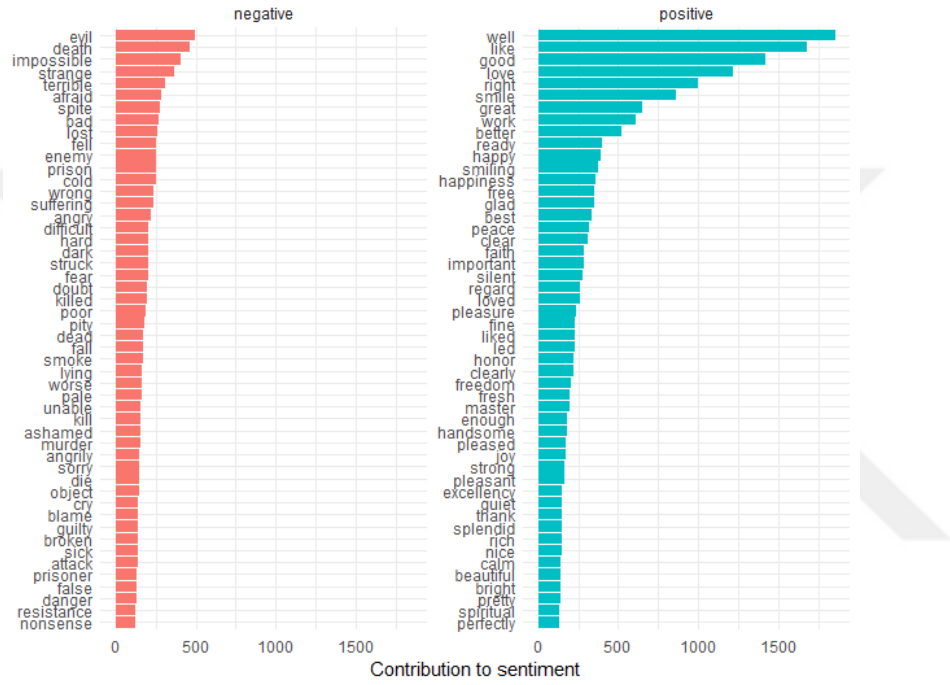
Tolstoy' un eserleri için yapılan duygu analizi çalışma sonuçları da aşağıda yer almaktadır. Çıkan grafikler incelendiğinde Katia (44266) adlı eser daha çok pozitif kelimelerden oluşurken My Religion (43794) adlı eser ise negatif kelimelerden oluşmaktadır. My Religion, dini inançlar ve eleştirileri ele almaktadır. Katia adlı roman 17 yaşındaki genç bir kadın ile kocası arasındaki aşkı anlatmaktadır. Bir aşk hikayesi olan romanda yapılan deneysel çalışma sonucunda pozitif kelimelerin çokluğu eserin mutlu bir aşk hikayesini anlattığını göstermektedir. Anna Karenina (1399) ise Rusların kendi ülkelerini ve dönemin aristokratlarını toplumsal ve insan psikolojisi yönü ile ele bir romandır. Romanda mutlu bir evliliğin evlilik dışı bir aşkın yol açtığı düş kırıklıklarını ve düşüşleri ele almaktadır. Bu sebeple yapılan duygu analizi frekans tablosunda da yer yer olumsuz yer yer de olumlu kelimelerin sıklığı dikkat çekmektedir.

Şekil 20: Tolstoy Eserleri İçin Duygu Analizi Frekans Tablosu



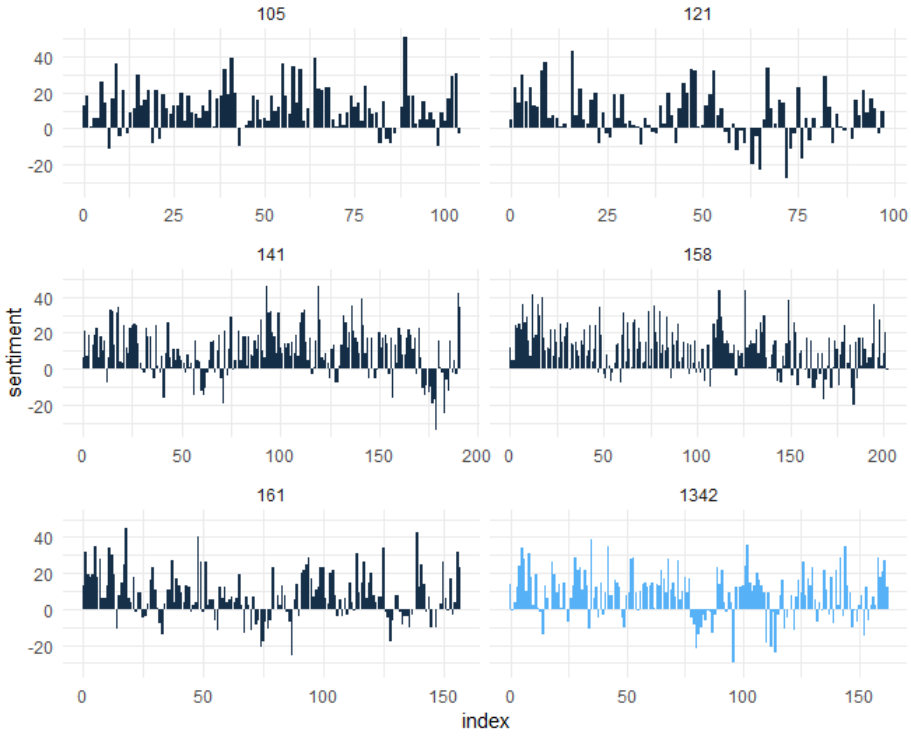
Şekil 20. incelendiğinde Tolstoy eserlerinde evil, death, die, prison, fall, bad, difficult gibi olumsuz kelimeler yer vermiştir. Bu da eserlerinde realist bir bakış açısı ile yazdığını daha çok savaş, din, mutsuzluk konularını ele aldığını göstermektedir.

Şekil 21: Tolstoy Eserleri İçin Positive Ve Negative Kelimeler



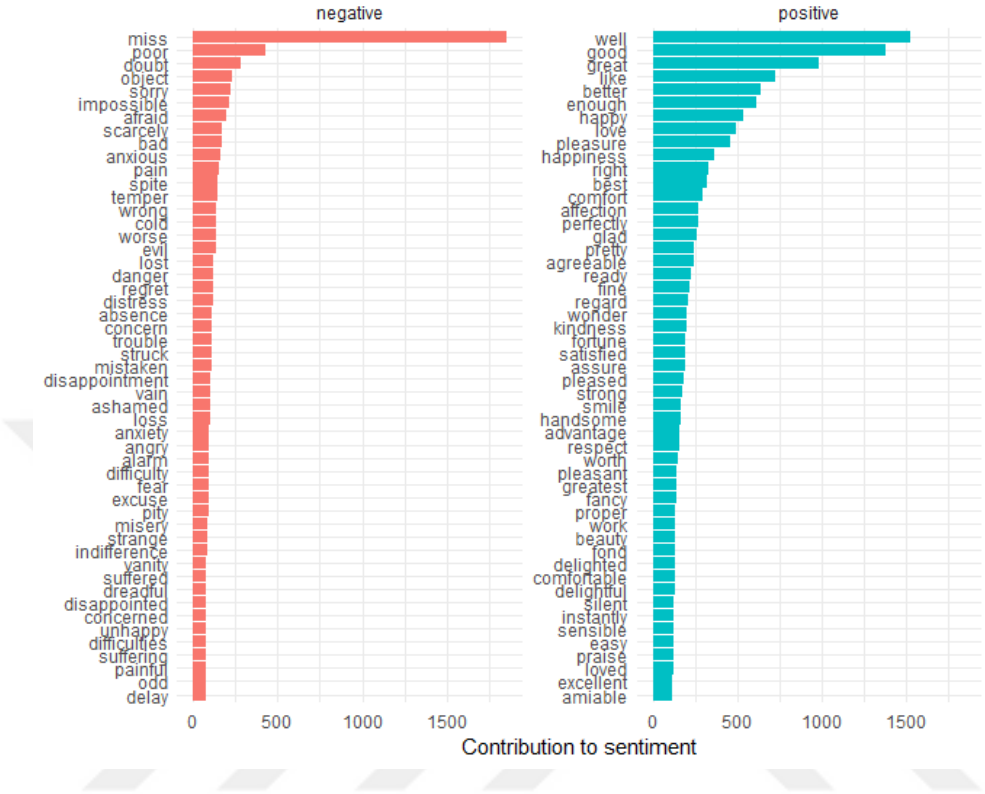
Jane Austen' un eserleri için yapılan duygu analizi çalışma sonuçları aşağıda Şekil 21' de yer almaktadır. Persuasion (105) adlı esere baktığımızda pozitif kelimelerin çoğunluğu dikkat çekmektedir. Persuasion adlı eser bir mutlu son ile biten bir aşk hikayesini anlatmaktadır. Yer yer hüzne yer yer sevinç ve mutluluğa yer vermektedir. Jane Austen duygu analizi frekans tablosu incelendiğinde eserlerinde olarak pozitif bir dil egemendir. Emma (158) romanında ise hem bir kasabada yaşayan üç genç kızın aşkı arayışını hem de insan yaradılışının zayıf yönlerini ve 19. yüzyıl İngiliz toplumunun sert geleneklerini alaycı bir dille ele almaktadır.

Şekil 22: Jane Austen Eserleri İçin Duygu Analizi Frekans Tablosu



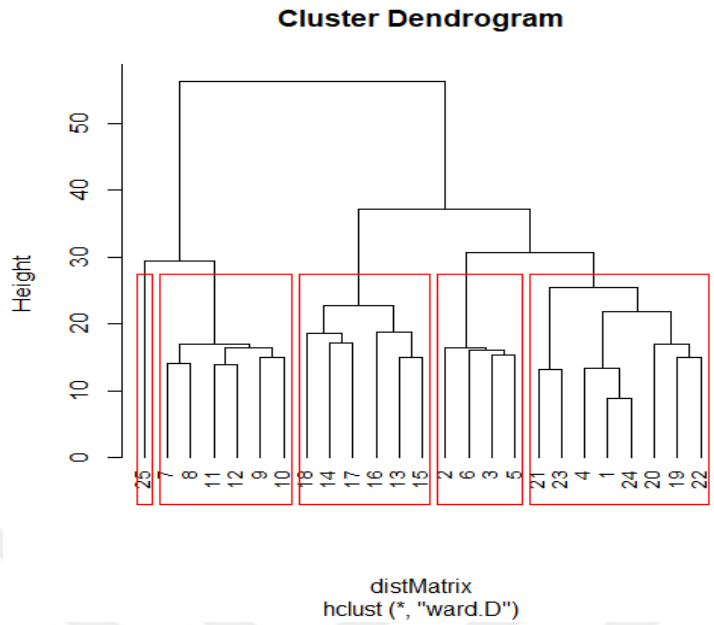
Şekil 22 incelendiğinde poor, pain, disappoinment gibi olumsuz kelimeler ile well, good, happy, love gibi olumlu kelimeleri sıklıkla kullandığı görülmektedir. Burada "miss" kelimesi negatif olarak kodlanmış olup Jane Austen "miss" kelimesini eserlerinde genç, evlenmemiş kadın anlamında sıfat olarak kullanmaktadır. Yapılan deneysel çalışma sonucunda Jane Austen için eserlerinde kadın kahramanları kullandığını, aşk ve mutluluk temalı olduğunu göstermektedir. Eserlerinde ne kadar hüzün ve mutsuzluk var ise de sonunla hep bir mutlu son olduğu sonucu çıkabilmektedir.

Şekil 23: Jane Austen Eserleri İçin Positive Ve Negative Kelimeler



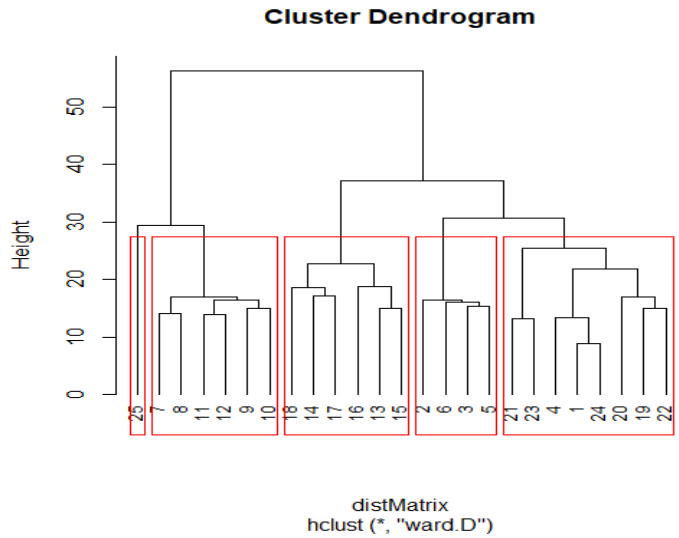
Deneysel çalışmamızın amacı olan yazarlık özellikleri belirlenerek yazarı belli olmayan bir eserin yazarının hangi kümeye ait olduğunun tespiti Jane Austen' ın Lady Susan adlı romanı ele alınarak Öklid uzaklık ölçüsü ve Ward kümeleme tekniği ile tespit edilmeye çalışılmıştır. Eserden yazar ile ilgili tüm bilgiler (Yazarın adı, soyadı, kitap ile ilgili özel isimler) R programı aracılığı ile kaldırılmıştır. 2. Adım olarak diğer eserlere de uygulanan metin ön işleme aşamaları uygulanmıştır. Eserin en sık tekrar eden kelimeleri, sıfatları ve zarfları bulunmuştur. Elde edilen sonuç aşağıda yer almaktadır. Aşağıdaki Ward kümeleme algoritmasının sonucu incelendiğinde 7 numara ile belirtilen Jane Austen' a ait olan ve çalışma kapsamında yazarı gizlenen Lady Susan adlı romandır. Dahil olduğu sınıfta (8, 9, 10, 11, 12) yer alan eserler ise sırasıyla Emma, ,Mansfield Park, Northanger Abbey, Persuasion, Pride and Prejudice adlı eserlerdir. Bu 5 eser de Jane Austen'a ait olan eserlerdir.

Şekil 24: Yeni eklenen eser ile yapılan kümeleme algoritması



Eserlerin yerleri değiştirilerek tekrar çalışma yapıldığında sonuç yine aynı çıkmaktadır. 1 numaralı eser Lady Susan iken, 20, 21, 22, 23, 24, 25 numaralı eserler ise Jane Austen' a ait olan diğer eserlerdir.

Şekil 25: Eserlerin yerleri değiştirilerek yapılan kümeleme algoritması sonuçları



SONUÇ

Bu çalışmada 4 farklı yazarın 6 farklı eseri ile ilgili yazarlık özellikleri çıkarılarak Öklid uzaklığı yardımıyla K-NN sınıflandırma algoritması kullanılarak yazarı bilinmeyen bir dokümanın önceden belirlenmiş 4 farklı yazar içerisinden hangisine ait olabileceği sorusunun cevabı aranmıştır. Yazarlık özellikleri elde edilirken İngilizce dili için çalışma yapılmıştır. İngilizce' de yer alan a, and, the gibi tek başına anlam ifade etmeyen kelimeler kaldırılmıştır. Tüm noktalama işaretleri kaldırılmış ve büyük harfler küçük harfe dönüştürülerek en sık tekrar eden kelimeler, sıfatlar ve zarflar bulunmuştur. Bu 3 özellik vektörünün sınıflandırma başarısı araştırılmıştır. Çalışmada K-NN kümeleme algoritması ve Ward kümeleme algoritması test edilmiştir. Her iki algoritmada da Öklid uzaklığı kullanılmıştır. KNN algoritmasının ve Ward algoritmasının veri seti üzerinde kullanımı için R dili ve R Studio aracı kullanılmıştır. Gerçekleştirilen deneysel çalışma sonucunda Öklid uzaklığı kullanılarak yapılan K-NN algoritmasının doğruluk oranı (ACC) %85 ve Ward kümeleme algoritmasının doğruluk oranı %75 olarak bulunmuştur.

Sonuç olarak yapılan deneysel çalışmada İngilizce metinler üzerinde bir metin madenciliği tekniklerinden yararlanılarak R programlama dili ile yazar tanıma ve sınıflandırma başarıları test edilmiştir. Metinler üzerinden duygu analizi yapılmıştır. Duygu analizi yardımıyla yazarın eserlerinde kullanmış olduğu dil bulunabilmektedir. Yazarın hangi türde eserler yazdığı tespit edilebilmektedir.

Tüm elde edilen bilgiler kullanılarak Jane Austen' a ait olan fakat tüm yazar bilgileri R programı aracılığı ile kaldırılan eser Ward kümeleme analizi ile sınıflandırılmış olup, 2 defa analiz edilmiştir. 2 kere farklı sıralama ile çalıştırılan kodlarda sonuç, yazarı belli olmayan eser Jane Austen' a ait olan diğer eserler ile aynı sınıfa dahil edilmiştir.

R programlama dili sahip olduğu paket ve kütüphaneler sayesinde İngilizce metinler üzerinde yapılacak olan metin madenciliği çalışmaları için oldukça uygundur. Yapılan çalışma sonucunda, R programında Türkçe metin analizi yapılabilmesi için Türkçe karakterlerin R programı yardımı ile analiz edilebilmesini sağlayan bir algoritma geliştirilebilir.

KAYNAKÇA

Aydın, G. (2017). *No-SQL Veri Tabanları Üzerinde Bir Metin Madenciliği Uygulaması* (Yüksek Lisans Tezi). İstanbul: İstanbul Aydın Üniversitesi Fen Bilimleri Enstitüsü

Aydilek, Berkan İ. (2013). *Veri Kümelerindeki Eksik Değerlerin Yeni Yaklaşımlar Kullanılarak Hesaplanması* (Yayınlanmamış Doktora Tezi); Konya: Selçuk Üniversitesi Fen Bilimleri Enstitüsü

Baykasoğlu, A. (2005). *Veri Madenciliği Ve Çimento Sektöründe Bir Uygulama*. Akademik Bilişim '05-VII. Akademik Bilişim Konferansı

Coşkun, C. ve Baykal, A. (2011). *Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri 2 - 4 Şubat 2011 İnönü Üniversitesi*, Malatya; Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması

Çelikyay, E. (2010). *Metin Madenciliği Yöntemiyle Türkçe 'de En Sık Kullanılan Ve Birbirini Takip Eden Harflerin Analizi Ve Birliktelik Kuralları* (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: Beykent Üniversitesi Fen Bilimleri Enstitüsü

Demiralay, M. ve Çamurcu, Y. (2005). "Cure, Agnes ve K-means Algoritmalarındaki Kümeleme Yeteneklerinin Karşılaştırılması", *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 8(2): 1-18

Dinler, M. (2014). *Kümeleme Analizi Yöntemlerinin Hayvancılık Verilerinde Karşılaştırılmalı Olarak İncelenmesi* (Yayınlanmamış Yüksek Lisans Tezi). Bingöl: Bingöl Üniversitesi Fen Bilimleri Enstitüsü

Doğan, S. (2006). *Türkçe Dokümanlar İçin N-gram Tabanlı Sınıflandırma: Yazar, Tür ve Cinsiyet* (Yayınlanmamış Yüksek Lisan Tezi). İstanbul: Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü

Dolgun ve Diğerleri (2009). Veri Madenciliğinde Yapısal Olmayan Verilerin Analizi: Metin ve Web Madenciliği, *İstatistikçiler Dergisi* 2, ss.48-58

Işık, M. ve Çamurcu Y. (2008) Web Belgeleri Kümelemede Benzerlik Ve Uzaklık Ölçütleri Başarılarının Karşılaştırılması. *Fen Bilimleri Enstitüsü Dergisi*, 20 Marmara Üniversitesi

Julia Silge and David Robinson (2016). Text Mining With R A Tidy Approach.

Kocamış, M. ve Durmaz, C. (2008). *Oracle Data Miner İle Öğrenci Kayıtları Üzerine Bir Veri Madenciliği Uygulaması* (Yayınlanmamış Bitirme Tezi). İstanbul: İstanbul Teknik Üniversitesi Fen Edebiyat Fakültesi

Köktürk, F. (2012). *K-En Yakın Nomşuluk, Yapay Sinir Ağları Ve Karar Ağaçları Yöntemlerinin Sınıflandırma Başarılarının Karşılaştırılması* (Yayınlanmamış Doktora Tezi) Zonguldak: Zonguldak Bülent Ecevit Üniversitesi Sağlık Bilimleri Enstitüsü

Amasyalı, M. F., Diri, B., ve Türkoğlu, F. (2006, June). *Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi*. In *Turkish Symposium on Artificial Intelligence and Neural Networks*.

Melek, C. (2012). *Metin Madenciliği Teknikleri İle Şirketlerin Vizyon İfadelerinin Analizi* (Yayınlanmamış Yüksek Lisans Tezi). İzmir: Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü

Nizam, H. ve Akın, S. (2014). *Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması*, XIX. Türkiye'de İnternet Konferansı.

Oğuz, B. (2009). *Metin Madenciliği Teknikleri Kullanılarak Kulak Burun Boğaz Hasta Bilgi Formlarının Analizi* (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: Akdeniz Üniversitesi Sağlık Bilimleri Enstitüsü

Özbay, E. (2007). *Finans Sektöründe Veri Madenciliği ile Dolandırıcılık Tespiti* (Yayınlanmamış Yüksek Lisans Tezi). Konya: Selçuk Üniversitesi Fen Bilimleri Enstitüsü

Pilavcılar, İ. F. (2007). *Metin Madenciliği İle Metin Sınıflandırma* (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü

Rouka, L. (2012). *Metinsel Veri Madenciliğinde Bilgisayarlı Çeviriciler* (Yayınlanmamış Yüksek Lisans Tezi). Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü

Sezgin, E. ve Çelik, Y. (2013) *Veri Madenciliğinde Kayıp Veriler İçin Kullanılan Yöntemlerin Karşılaştırılması; Akademik Bilişim 2013 – xv*. Akademik Bilişim Konferansı

Şeker, Sadi E. (2015) *YBS Ansiklopedisi*. 2(4).

Şeker, Sadi E. ve Eşmekaya, E. (2017). *Eksik Verilerin Tamamlanması* (IMPUTATİON), *Yönetim Bilişm Sistemleri*. Ansiklopedi. 4(3).

Şengöz, N. ve Özdemir, G. (2016). Temel Bileşenler Analizi Ve Ve K-Ortalama Kümeleme Yönteminin Birlikte Kullanımı: Bir Örnek Uygulama-Combined Use Of Principal Component Analysis And K-Clustering Method: A Case Study. Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 8(15), 85-94.

Şengür, D. ve Tekin A. (2013). Öğrencilerin Mezuniyet Notlarının Veri Madenciliği Metotları İle Tahmini. *Bilişim Teknolojileri Dergisi*. 6(3).

Taşova, O. (2011). Yapay Sinir Ağları İle Yüz Tanıma (Yayınlanmamış Yüksek Lisans Tezi). İzmir: Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü

Temel, G. ve Erdoğan, S. (2011). Tanı Koyma Amaçlı Yapılan Tıbbi Çalışmalarda Jacknife, Bootstrap Ve Çapraz Geçerlilik Yöntemlerinin Kullanımı. Düzce Üniversitesi. Sağlık Bilimleri Enstitüsü Dergisi. 1(3): 45-49

Tetik, S. (2016) Text Mining Application with Twitter and R (Lisans Bitirme Tezi). İstanbul: Yıldız Teknik Üniversitesi Fen-Edebiyat Fakültesi İstatistik Bölümü

Tiryaki, S. (2006). *Lojistik Alanında Bir Veri Madenciliği Uygulaması* (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

U, İlhan. (2001) *Application Of KNN and FPTC Based Text Categorization Algorithms to Turkish News Reports* (Yayınlanmamış Doktora Tezi). Bilkent Üniversitesi Fen Bilimleri Enstitüsü

Yapay Sinir Ağları Sempozyumu, Muğla, 21- 24 Haziran, 2006.

Yıldız, K. , Çamurcu, Y. ve Doğan, B. (2010). *Veri Madenciliğinde Temel Bileşenler Analizi Ve Negatifsiz Matris Çarpanlarına Ayırma Tekniklerinin Karşılaştırmalı Analizi Akademik Bilişim - XII. Akademik Bilişim Konferansı Bildirileri*. Muğla Üniversitesi.

Yılmaz, Ş. ve Patır, S. (2011). Kümeleme Analizi Ve Pazarlamada Kullanımı; *Akademik Yaklaşımlar Dergisi*. 2(1):1

Silge, J. And Robinson, D. (2013). Text Mining with R: A Tidy Approach

EKLER

Ek 1: Yazar-Eser Listesi

1 = Martin Guerre (Alexandre Dumas)

2 = Ten Years Later (Alexandre Dumas)

3 = The Black Tulip (Alexandre Dumas)

4 = The Count Of Monte Cristo (Alexandre Dumas)

5 = The Man İn The Iron Mask (Alexandre Dumas)

6 = The Three Musketeers (Alexandre Dumas)

7 = Emma (Jane Austen)

8= Mansfield Park (Jane Austen)

9 = Northanger Abbey (Jane Austen)

10 = Persuasion (Jane Austen)

11 = Pride And Prejudice (Jane Austen)

12 = Sense And Sensibility (Jane Austen)

13 = A Tramp Abroad (Mark Twain)

14 = Adventures Of Huckleberry (Mark Twain)

15 = Life On The Mississippi (Mark Twain)

16 = Roughing İt (Mark Twain)

17 = The Adventures of Tom Sawyer (Mark Twain)

18 = The Prince And The Pauper (Mark Twain)

19 = Anna Karenina (Tolstoy)

20 = Katia (Tolstoy)

21 = My Religion (Tolstoy)

22 = Resurrection (Tolstoy)

23 = The Kingdom Of God İs Within You (Tolstoy)

24 = War And Peace (Tolstoy)

Ek 2: Eserler ve Gutenberg Kodları

Alexandre Dumas

The Man in the Iron Mask (2759)

Martin Guerre (2752)

Ten Years Later (2681)

The Count of Monte Cristo (1184)

The Three Musketeers (1257)

The Black Tulip (965)

Mark Twain

Roughing It (3177)

Life on the Mississippi (245)

The Adventures of Tom Sawyer (74)

Adventures of Huckleberry Finn (76)

The Prince and The Pauper (1837)

A Tramp Abroad (119)

Leo Tolstoy

Anna Karenina (1399)

My Religion (43794)

Resurrection (1938)

Katia (44266)

The Kingdom of God is Within You (43302)

War and Peace (2600)

Jane Austen

Emma (158)

Mansfield Park (141)

Northanger Abbey (121)

Persuasion (105)

Pride and Prejudice (1342)

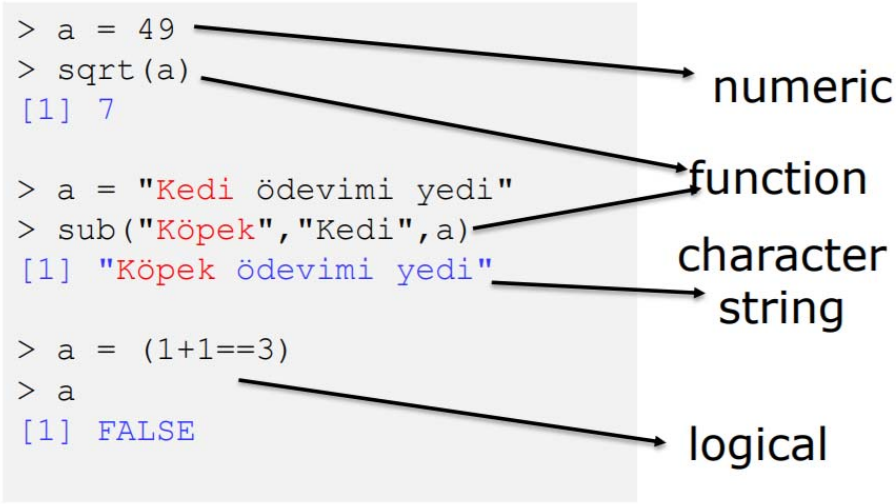
Sense and Sensibility (161)

Ek 3: R Programlama Dili

R Programındaki Temel Nesne Türleri

- Numeric
- Character
- Logical
- Function

Bu nesneler kullanılarak elde edilen objeler aşağıda gösterilmiştir.



Kaynak: Baydoğan, Çetin ve Orbay, R, inet-tr'14, İzmir

Temel Komutlar

R programı ile ilgili tüm kodlara ve bu kodların detaylı açıklamalarına https://cran.r-project.org adresinden ulaşılabilmektedir.

getwd() = R içerisinde temel çalışma dizinini gösterir.

setwd() = Çalışma dizinini değiştirir.

ls() = Tüm nesneleri listeler.

ls.str() = Tüm nesnelerin detaylarını listeler.

str(data) = <data>'in detaylarını listeler.

list.files() = Dizindeki tüm dosyaları listeler.

dir() = Dizindeki tüm dosyaları listeler.

nesnem = Basit olarak nesneyi görüntüler.

rm(nesnem) = <nesnem> nesnesini kaldırır.

rm(list=ls()) = Çalışma alanındaki tüm nesneleri kaldırır.

save(dosya, file="text.rda") = <dosya> nesnesini <text.rda> dosyasına kaydeder.

load("text.rda") = "text.rda" yı hafızaya yükler.

summary(text) = "text" için basit istatistikleri görüntüler.

hist(x,freq=TRUE) = <x> nesnesine ait histogramı görüntüler. <freq=TRUE>

frekansı gösterir ve <freq=FALSE> olasılıkları gösterir.

q() = R' dan çıkar.

Text Dosyasının R Programına Okutulması

kitap.v <- scan ("book.txt", what="character", sep="/n")

Burada **kitap.v** kullanıcının belirlediği ve bundan sonraki işlemleri bu isim ile yapacağı bir vektördür.

Scan = Text dosyasını R programına yazdırılmasını sağlar.

Sep = Her bir satırın vektörde ayrı bir veri nesnesi olmasını sağlamaktadır.

Csv. Uzantılı Dosyaların R Programına Okutulması

dosya <-read.csv(file, header = TRUE, sep = ",", dec = ".", ...)

file = dosyanın ismi (eğer farklı bir konumda ise tam yolu ile birlikte, örneğin "c:\r verileri\örnek.csv")

header = veri setinde değişken isimleri yer alıyor (True/False)

sep = veri noktalarını ayıran karakter (“; , .”)

dec = ondalık ayracı için kullanılan karakter

ornek<- read.table(“C:\\Users\\eda çam\\Documents\\data\\dumas\\Ten Years Later1.csv” , header=T,sep=”,”)

İstatistik paketlerinin Yüklenmesi

Install.package (“tm”) = Text Mining paketinin yüklenmesini sağlar.

Burada dikkat edilmesi gereken kısım ilgili paket isminin hem parantez hem de tırnak içinde yazılı olmasıdır.

library (tm) = Yüklenen ***“tm”*** paketinin kütüphaneden çağrılmasını sağlar.

R Programı İle Metin Analizi

Library(dplyr) = Verilerin düzenlenmesi, filtrelenmesi, sıralanması ve belirlenen değişkenlerin hesaplanmasını sağlayan bir komuttur. ***Select(), mutate(), filter(), arrange(), summarize()*** ve ***group_by()*** gibi fonksiyona sahiptir.

Library(tidytext) = Metni düzenli veri çerçevesinde yazdırmayı sağlar. Kütüphane içerisinde bulunan ***“unnest_token”*** fonksiyonu sayesinden metni kelime kelime ayırıp analiz edilmesine olanak sağlamaktadır.

Data(stop_words) = Durdurma kelimeler yüklenir. ***“Anti_join”*** fonksiyonu ile metin içerisinden kaldırılır.

Library(ggplot2) = Verilerin grafik ile gösterilmesini sağlamaktadır.

Sentiment = Sentiment fonksiyonu “***tidytext”*** kütüphanesi içerisinde yer alır.

Sentiment fonksiyonunda 3 adet (NRC, BİNG ve AFINN) sözlüğü bulunmaktadır. Bu sözlükler duygu analizinin yapılabilmesine yardımcı olmaktadırlar.

İnner_join = Duygu analizi için içsel birleştirme fonksiyonudur. Aşağıdaki örneğe bakılırsa "***Filter"*** ile analiz edilecek kitap seçilmiş olup "***inner_join*** "ile ***nrc*** kütüphanesinde olumlu, mutlu anlam taşıyan kelimeler "***count"*** fonksiyonu ile kelime kelime listelenmiştir. "Sort = TRUE" foksiyonu ile kelimeler tekrar sırasına göre sıralanır.

```
books %>%
 filter(book == "KIŞ") %>%
inner_join(nrc_joy) %>%
count(word, sort = TRUE)
```

Top_n = İlk 50 veya ilk 10 gibi sıralamaların yapılasını sağlamaktadır. Örnek olarak; **"top_n (20)"** komutu yazdırılırsa ilk 20 kelimeyi listelenir.4

library(wordcloud) = Kelime bulutunun oluşturulmasını sağlar. En sık tekrar eden kelimeler, pozitif, negatif kelimeler gibi filtre bazında kelime bulutu oluşturulur.

```
Veri_seti %>%
anti_join(stop_words) %>%
count(word) %>%
 with(wordcloud(word, n, max.words = 100))
        # Burada veri seti içerisindeki stop
word' ler kaldırıldıktan sonra en sık tekrar
eden ilk 100 kelime kelime bulutu olarak
yazdırılmıştır.
```

Kaynak: Text Mining with R : A Tidy Approach ; Julia Silge, David Robinson

R Programında Metin İçerisindeki Sıfat Ve Zarfların Bulunulması

Library(udpipe) = Metin madenciliği için Doğal Dil İşleme (NLP) analizinin yapılmasını sağlamaktadır.

Doğal Dil İşleme (NLP) = Doğal dil işleme, doğal dilde metinlerin ve/veya seslerin içerisinden yararlı ve anlamlı bilgilerin bilgisayar ortamında işlenerek bilgisayarın anlayabileceği bir yapıya çeviren, bilgisayar bilimi ve dil bilimi ile ilişkili bir bilim dalıdır.

Library(textrank) = Metnin özetlenmesini sağlar.

Library(RColorBrewed) = Grafikler için renk paletlerinin yüklenmesini sağlar.

TermDocumentMatrix( ) = "tm" paketi içerisinde yer alan Terim – Doküman Matrisinin oluşturulmasını sağlar.

Cluster( ) = İçerisinde tüm veri kümeleme tekniklerini bulunduran fonksiyondur. (Hiyerarşik kümeleme, K-En yakın komşu algoritması, Ward vb.)

```
cluster  <- as.matrix(tdm.stack.nl)
 distMatrix <- dist(m, method="euclidean")
 groups <- hclust(distMatrix,method="ward.D")
 plot(groups, cex=0.9, hang=-1)
  rect.hclust(groups, k=5)
```

Yukarıdaki örnekte uzaklık ölçüsü olarak Öklid uzaklık ölçüsü seçilmiş olup, kümeleme tekniği olarak ise Ward tekniği seçilmiştir. Küme sayısı 5 olarakalınmıştır.

Library(philentropy) = 46 farklı benzerlik ve uzaklık ölçülerini içermektedir.