KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

METİNSEL VERİ MADENCİLİĞİNDE BİLGİSAYARLI ÇEVİRİCİLER

YÜKSEK LİSANS TEZİ

Bilgisayar. Müh. Leila ROUKA

HAZIRAN 2012

TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

METİNSEL VERİ MADENCİLİĞİNDE BİLGİSAYARLI ÇEVİRİCİLER

Leila ROUKA

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"BİLGİSAYAR YÜKSEK MÜHENDİSİ"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih : 25.05.2012
Tezin Savunma Tarihi : 20.06.2012

Tez Danışmanı : Doç. Dr. Cemal KÖSE

Trabzon 2012

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalında

Leila ROUKA tarafından hazırlanan

Metinsel Veri Madenciliğinde Bilgisayarlı Çeviriciler

başlıklı bu çalışma, Enstitü Yönetim Kurulunun 20 / 06 / 2012 gün ve 1350 sayılı kararıyla oluşturulan jüri tarafından yapılan sınavda

YÜKSEK LİSANS TEZİ

olarak kabul edilmiştir.

Jüri Üyeleri

Başkan : Doç. Dr. Cemal KÖSE …...……

Üye : Doç. Dr. Ali GANGAL …...……

Üye : Yrd. Doç. Dr. Hüseyin PEHLİVAN ……...……

Prof. Dr. Sadettin KORKMAZ

Enstitü Müdürü

ÖNSÖZ

Güncel ve faydalı bu tez konusu seçiminde bana yol gösteren, benden hiç bir zaman yardımını esirgemeyen Sayın Danışman Hocam Doç. Dr. Cemal KÖSE' ye, yardım ve desteğinden dolayı sonsuz teşekkür ve şükranlarımı sunarım.

Tez çalışmalarımda desteklerini esirgemeyen Bilgisayar Mühendisliği bölümündeki hocalarıma teşekkür ederim.

Tez çalışmalarımda beraber çalıştığım ve tezdeki en zor zamanlarımda hep yanımda olan arkadaşım Parham TOFİGHİ' ye yardımları için sonsuz teşekkürlerimi ve saygılarımı sunarım.

Öncelikle, beni yetiştirip bu günlere getiren sevgili Annem Jaleh GAFFARİ ve Babam Gazanfar ROUKA, ve tüm aileme saygı ve sevgilerimi sunarım.

Ayrıca, çalışmaya emeği geçen, ismini yazamadığım tüm arkadaşlarıma ve Karadeniz Teknik Üniversitesi' ne teşekkürlerimi sunarım.

Leila ROUKA

Trabzon 2012

TEZ BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum "Metinsel veri madenciliğinde bilgisayarlı çeviriciler" başlıklı bu çalışmayı baştan sona kadar danışmanım Doç. Dr. Cemal KÖSE 'nin sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deneyleri/analizleri ilgili laboratuarlarda yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 25/05/2012

Leila ROUKA

İÇİNDEKİLER

Yüksek Lisans

ÖZET

METİNSEL VERİ MADENCİLİĞİNDE BİLGİSAYARLI ÇEVİRİCİLER

Leila ROUKA

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Doç. Dr. Cemal KÖSE
2012, 66 Sayfa

İnternetteki metinsel bilgilerin büyümesi, istenilen bilgiye etkin biçimde erişimini gittikçe zorlaştırmaktadır. Metinsel veri sınıflandırma yöntemleri bu soruna etkili çözümler sunmaktadır. Metinsel veri sınıflandırma, bir takım belgeleri önceden tanımlanmış kategoriler içinde otomatik olarak sıralama görevidir. Son yıllarda bu konuda pek çok dillerde çalışan birçok araştırmacı birçok farklı araştırma ve geliştirme çalışmaları yapmaktadır. Fakat, bu çalışmalar çoğunlukla orijinal metinler üzerinde yapılmaktadır. Bu çalışmada, metinsel veri sınıflandırmada bilgisayarlı çeviricilerin etkisi değişik sınıflandırma yöntemleri kullanılarak incelenmiştir. Geliştirilen sistem ilk olarak orijinal dildeki metini analiz edip sınıflandırmakta ve daha sonra aynı metini bilgisayarlı çeviriciler kullanarak hedef dile çevirmekte ve çevrilen metini orijinal dildeki gibi aynen analiz ederek sınıflandırmaktadır. Daha sonra, metinsel veri sınıflandırmada bilgisayarlı çeviricilerin etkisini ölçmek için elde edilen sonuçlar karşılaştırılmıştır. Bu çalışmada, kullanılan sınıflandırma yöntemi performansları da ölçülmüş ve karşılaştırılmıştır. Elde edilen sonuçlara göre Multinomial Naive Bayes yöntemi en başarılı yöntemdir. Yine, aynı belgenin farklı dillere çevrilmiş sınıflandırma sonuçları dikkate alındığında, bilgisayarlı çeviricilerin metinsel veri sınıflandırmada oldukça az bir etkisi olduğu görülmüştür. Bu sonuçlar bilgisayarlı çeviricilerin bir dil temel alınarak farklı dillerde veri madenciliğinin oldukça etkin bir şekilde yapılabileceğini göstermektedir.

Anahtar Kelimeler: metin madenciliği, çevrilen metinler, sınıflandırma algoritmalar, bilgisayarlı çeviriciler.

Master Thesis

SUMMARY

MACHINE TRANSLATOR IN THE TEXTUAL DATA MINING

Leila ROUKA

Karadeniz Technical University
The Graduate School of Science
Computer Engineering Graduate Program
Supervisor: Assoc. Prof. Dr. Cemal KÖSE
2012, 66 Pages

With the growth of online textual information, effective information access is difficult without good classification and summarization of document content. The textual data classification methods offer efficient solutions to this problem. Defined text classification (or categorization) automatically sort documents in a number of predefined categories. In recent years, many researchers working on this issue has conducted many research studies in different languages but most of these studies are carried out on English texts. In this study, we evaluate the efficiency of Machine Translators on the Web-based texts classification, by classification Original texts into predefined categories and then translating them into other language with machine translator for accomplish classification operation with the same categories and assess the results in two situations. In addition to this, the effect of machine translators in the textual data classification is examined by using supervised classification methods. The developed system first analyzes and classifies an input text in one language, and then analyzes and classifies the same text in another language generated by machine translators from the input text. The obtained results are compared to measure the effect of the translators in textual data classification. The performances of the classification method used in this study are also measured and compared. The obtained results show that Multinomial Naïve Bayes method is the most successful method, and that the machine translation has quite a small effect on the attained classification accuracy.

K**eywords:** Text Mining, Translated Text, Classification Algorithms, Machine Translator.

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

SEMBOLLER DİZİNİ

VTBK	: Veri tabanlarında bilgi keşfi
SVM	: Support Vector Machines
NB	: Naïve Bayes
MNB	: Multinominal Naïve Bayes
SMO	: Sequential Minimal Optimization
NLP	: Natural Language Processing
CRM	: Customer Relationship Management
XML	: Extensible Markup Language
HTML	: Hyper Text Markup Language
IG	: Information Gain
TF	: Term Frequency
IDF	: Inverse Document Frequency
Tfc	: Term Frequency Component
Ltc	: Logarithmic Term Component
TP	: True Pozitif
FN	: False Negatif
FP	: False Pozitif
TN	: True Negatif
CSV	: Comma-Separated Values
CV	: Cross Validation
PS	: Percentage Split
İng	: İngilizce
Trk	: Türkçe

1. GENEL BİLGİLER

1.1. Giriş

Çok büyük miktardaki stratejik bilgiler, günümüzde ilerleyen bilgi teknolojileri sayesinde çeşitli şekillerde yayılmaktadır. Yine her geçen gün, teknolojinin gelişmesiyle veri miktarı çeşitli alanlarda hızla artmaktadır. Örneğin:

- İş dünyasında**:** E-ticaret ve ticari rekabet baskısının artması, Webdeki alışveriş ve hisse senetlerinin genişlemesiyle bilgi miktarı her geçen gün artmaktadır. Böylece, iş dünyasında karar vermek için bu bilgilerin içinden anlamlı bilgilerin erişimi ve dönüşümü, küresel rekabet gücünün karşısında çok önemli bir esas haline gelmiştir.
- Bilim dünyasında: Bioinformatik, simülasyonlar, uzaktan algılama ve izleme.
- Diğer alanlarda: Haberler, sayısal kameralar, youtube, facebook, kişisel notlar, CSR yönetimi ve benzerleri.

Bilgiyi üreten ve kullanan sorumlu yerler hızla gelişip sürekli ilerlediği için bilginin daha etkin bir şekilde yönetilmesi gerekir. 1990'ların ortalarından beri, bu kaynaklardan yararlı bilginin keşfedilmesini destekleyen teknikler, yöntemler ve araçların oluşturulması için pek çok araştırma yapılmıştır [6]. Bilgi çağında, değer oluşturmak fiziksel varlıklardan ziyade bilgi kaynaklarını etkin bir şekilde kullanmaktan geçiyor. Bu amaçla, bilgi yönetimi için birçok yöntem ve teknik, geliştirilmektedir. Veri madenciliği de bu amaca ulaşmak için kullanılan bir tekniktir. O, anlamsız veriden anlamlı bilgileri seçerek veriler arasındaki ilişkileri belirlemektedir. Genel anlamıyla veri madenciliği, verileri analiz etme, veriler arasından anlamlı bilgiye ulaşma ve ulaşılan bu bilgileri yararlı olacak şekilde özetleme işidir [30]. Çeşitli alanlarda kullanılan veri madenciliğinin Web madenciliği ve metin madenciliği gibi iki alt alanı vardır.

Herhangi bir dilde yazılmış Web metinleri üzerinde birçok veri madenciliği çalışmaları vardır. İnternet ortamında mevcut bu devasa metin bilgileri iyi bir şekilde

indekslenirse, yararlı bilgilere daha etkin şekilde ulaşılması kolaylaşır. Web ve metin madenciliği yöntemleri, bu soruna etkili çözümler sunmaktadır. Bu çalışmada, temel olarak, bu yöntemlerin bilgisayarlı çevirici ile elde edilmiş metinler üzerinde nasıl bir etki göstereceğinin analizi hedeflenmektedir.

Metin sınıflandırma yöntemi, birçok belgeyi önceden tanımlanmış kategoriler içinde otomatik olarak sınıflandırma işlemi olarak tanımlanabilir. Metinsel veri sınıflandırmada kullanılan Bayes, karar ağaçları ve mesafe tabanlı algoritmalar gibi çeşitli yöntemler vardır [27].

İngilizcenin İnternette ve dünya genelinde kullanılan en yaygın dil olmasından dolayı, metinsel veri sınıflandırma da genellikle İngilizcede gerçekleştirilir. Bununla beraber diğer dillerde de yapılan birçok çalışmalar vardır. Diğer taraftan, bu dillerin yapıları birbirinden çok farklıdır. Örneğin; İngilizce 26 harfli ve morfolojik yapısı nispeten basit bir dildir. Buna karşılık Türkçede 29 harf bulunur ve Türkçe nispeten daha karmaşık morfolojik bir yapıya sahiptir [9], [11]. Ancak bu güne kadar yapılan çalışmalar, orijinal dillerde yazılmış metinler üzerinde değerlendirilmektedir. Bu çalışmada, ilk olarak orijinal İngilizce ve Türkçe yazılı belgeler üzerinde aynı sınıflandırma algoritmaları kullanılarak sınıflandırılmış, daha sonra belgeleri bilgisayarlı çevirici ile çevirerek tekrar aynı sınıflandırma algoritmaları kullanılmıştır. Böylece, orijinal dildeki metinler ve çevirilerden elde edilmiş dokümanları ayrı ayrı sınıflandırarak, bilgisayarlı çeviricilerin sınıflandırma üzerindeki etkisi ölçülmüş ve değerlendirilmiştir. Sonraki bölümlerde veri madenciliğinin tanımı ve alt alanları, metin sınıflandırma süreci, yapılan çalışmalar, bu çalışmaların sonuçları ve uygulamalar ele alınmıştır.

1.1. Veri Madenciliği

Dijital saklama ambarlarında veri miktar büyümesi ve bilgisayar sistemlerinin kullanım artışları, veri depolarının büyük boyutlara ulaşmasına neden olmuştur. Veri madenciliği, bilgisayar programı kullanarak büyük hacimdeki verilerin farklı yöntemler ile analiz edilmesini ve anlamlı bilgilerin çıkartılması veya tahminde bulunabilmeyi sağlayacak ilişkileri arar [8], [10]. Şekil 1.de veri madenciliğinin bu durumu ifadesi gösterilmektedir [12].

Şekil 1. Veri Madenciliği [12].

Veri Madenciliği için diğer tanımlar aşağıda özetlemiştir:

Veri madenciliği önceden bilinmeyen ve potansiyel olarak faydalı olabileceği düşünülen verilerin içerisindeki gizli bilgilerin çıkarılmasına denir. Diğer bir açıdan da veri madenciliği, büyük veri kümesi içinde depolanan genel ilişkilerin ve örüntülerin çıkarılması olarak verilebilir.

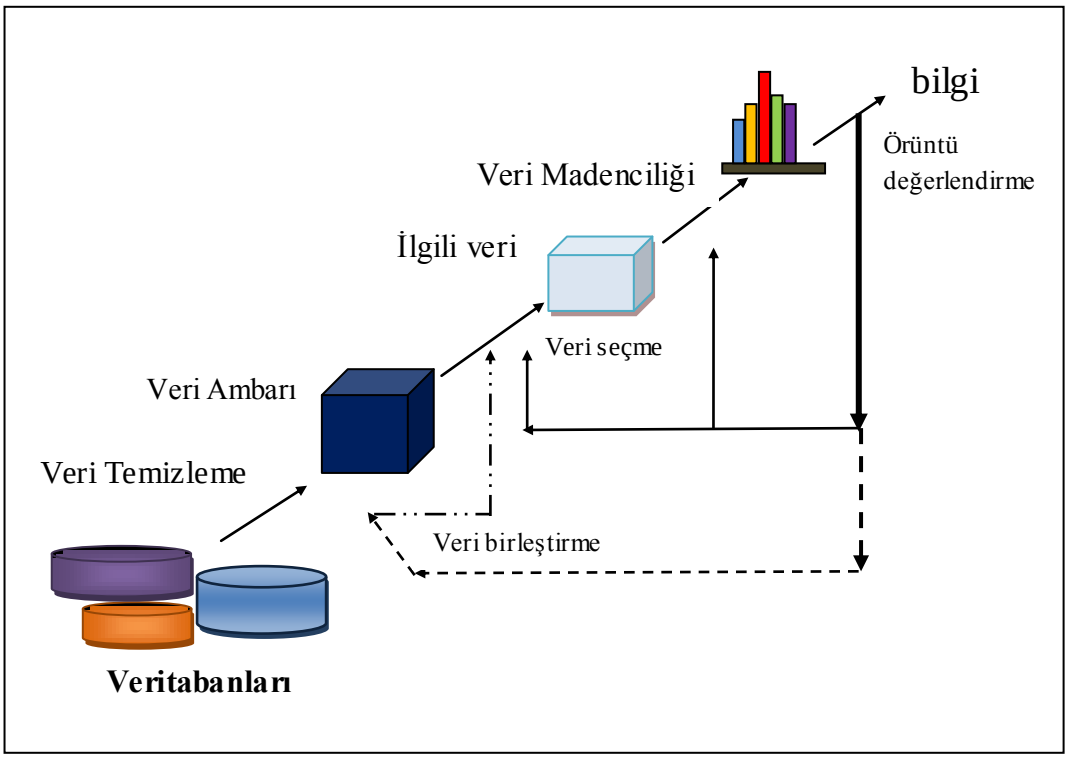
Bazı tanımlamalara göre, veri madenciliği, veri tabanlarında bilgi keşfi, büyük hacimli veri kümelerini tamamen veya yarı otomatik olarak analiz eden yöntem ve tekniklerin geliştirilmesi ve araştırılması olarak değerlendirilmektedir [2]. Veri madenciliği, veri tabanlarında bilgi keşfinin aşamalarından biridir. Buna göre VTBK (Veri Tabanlarında Bilgi Keşfi), veri temizleme, veri birleştirme, veri seçimi, veri indirgeme, veri madenciliği ve değerlendirme basamaklarından oluşan bir işlemdir. Böylece, veri madenciliği veri depolarında analiz edilmeyen ancak anlamlı ve yararlı desenleri, büyük miktarda olan veritabanlarından otomatik biçimde elde edilmesini sağlayan VTBK süreci içinde bir adımdır [23].

Veri madenciliği, istatistik alanındaki pek çok metodu kullanmasına rağmen, nesnelerin özelliklerine ve değerlerine bağlı sonuç vermede bilinen istatistiksel metotlardan ayrılmaktadır. Veri madenciliği disiplini oluşmadan önce istatistiksel teknikler, karar alma mekanizmasında sık sık kullanılmaktaydı. Ancak, kullanılan bu yöntemlerin sorunu, veri madenciliği algoritmalarının uygulama kolaylığı ile kıyaslamada, veri nedenleme sürecindeki en güç adımı oluşturuyordu [12], [23].

1.1.1. Veri Madenciliğinin Aşamaları

Veri madenciliğinin temelinde, veri önişleme, verinin analiz edilmesi, veri ambarlarının benzerliklerinin ve ilişkilerinin çıkarılması için bilgisayar programlama teknikleri ve istatistiksel metotların uygulaması gibi işlemler vardır. Anlamsız verilerden anlamlı bilginin ortaya çıkarılması için veriler birçok işleme tabi tutulurlar ki bu işlemler veri madenciliğinin aşamalarını oluşturur. Veri önişleme, çok fazla verinin bulunduğu veritabanı veya veri ambarlarındaki verileri analiz etmeden önce istatistiksel olarak sağlıklı hale getirmeyi amaçlamaktadır. Veriyi kullanılabilir hale getirmek için eksik, yetersiz, tutarsız, aykırı özellik taşıyanların belirlenip uygun yöntemlerle bunlara çözüm bulunması gerekmektedir [25].

Genellikle, pek çok araştırmacı tarafından veri madenciliği ve veritabanlarında özbilgi keşfi (Knowledge discovery in databases) aynı anlamda kullanılır. Oysa veri madenciliği, bilgi keşfi işleminin bir parçasıdır ve onun basamaklarından birisi sayılır. Veritabanlarında özbilgi keşfinin adımları özet olarak Şekil 2'de verilmiştir [2].



Şekil 2. Veritabanlarındaki özbilgi keşfinin aşamaları [2].

Şekil 2'de verildiği gibi veritabanlarında özbilgi keşfi aşağıdaki aşamalardan oluşur:

1-Veri hazırlama (Veri Temizleme, Veri Birleştirme ve Veri Seçme): Verilerin bilgisayar kullanımı ve işleme için uygun şekle çevrilmesidir.

2-Veri madenciliği: Veri ambarlarının içindeki örüntüleri, ilişkileri, düzensizlikleri, değişiklikleri, yönetimleri ve istatistiksel yapılarını keşfetmek için önemli metotları ve algoritmaları kullanmaktır.

3-Örüntü değerlendirme: Bilgileri temsil eden bazı ölçümlere göre uygun örüntüleri tanımlanması ve modellemesidir.

4-Bilgi sunumu: Madenciliği gerçekleşmiş ve tanımlanmış bilginin kullanıcılara rapor edilmesidir.

Bu aşamalarda, ham veri işlenir, yapısal veri elde edilir, veriler arası örüntüler ortaya konulur, bu örüntüler modellenir ve sonuçta bilgi keşfi sağlanır [10], [16].

Büyük boyutlu verilerin içindeki bilgiye erişme, diğer bir ifadeyle büyük veri kümeleri içerisinden, bilgisayar programı kullanarak gelecekle ilgili tahminde bulunabilmemizi sağlayabilecek bağıntıların aranması, veri madenciliğinin hedeflerinden biridir. Haber sitelerinde bulunan haberlerin daha çok hangi kategoriye ait olduğunun tespit edilebilmesi örnek olarak verilebilir. Veri madenciliğinde değişik örüntüler kullanıcıya açıklanır ve bunlar gerekirse bilgi tabanına da kaydedilebilir. Böylece, veri madenciliği süreci, gizli örüntülerin bulunmasına kadar devam eder [16]. Diğer bir açıdan, veri madenciliği, veri kümelerinin arasındaki desenlerin, veri analiz ve yazılım metotlarının kullanılmasıyla ilişkilendirilir. Burarda veri, her çeşit sayısal veya mantıksal değerdir. Öznitelik ise bir nesneye dayanan özellik ve onu tarif eden bir değerdir veya onun karakteristiğidir. Öznitelik değerleri, bir özniteliğe karşı gelen sayılar veya sembollerden ibarettir. Özniteliklerin çeşitli türleri (nominal, ordinal, interval ve ratio) vardır ve bunlar ölçüm seviyelerini ifade ederler [7], [30]. Veriler içindeki bağlantıların, özelliklerin, kuralların ve ilişkilerin bulunmasından bilgisayar sorumludur [30].

1.1.2. Veri Madenciliğinin Önişlemleri

1.1.2.1. Veri Tanımlama ve Özetleme

Veri açıklama ve özetlemenin maksadı, veri niteliklerinin sade bir şekilde az ve öz olarak açıklanmasıdır. Yani veri yapısının tanımlanmasıdır. Böylece, veri açıklama ve özetleme bir veri madenciliği işleminin bir planı olabilir. Veri tanımlama ve özetleme çoğunlukla diğer veri madenciliği problemleriyle bir arada uygulanır. Özetleme bulunan sonuçların hazırlamasında önemli bir etkiye sahiptir. Veri madenciliğinde farklı konu türlerinin sonuçları, daha üst seviyedeki bir verinin özetlenmesi olarak düşünülebilir.

Veri madenciliği başlanıldığı zaman veri analizinin niyeti ve verinin niteliği tamamen tespit edilemeyebilir. Basit betimsel istatistikler ve görselleştirme yöntemleri kullanılarak keşifsel veri analizleri yapılır ve bu metotlar, verinin niteliğinin anlaşılmasını ve gizli bilgilerin bulunmasını sağlar.

1.1.2.2. Veri Madenciliğinde Veri Hazırlama

Veri, çeşitli şekillerde ortaya çıkabilir. Veri madenciliği sisteminin kullandığı sayısal veya mantıksal veriler ise her türlü özellik veya karaktere sahip olan bir değerdir [21]. Kullanılacak veriler, yapısal veya yapısal olmayan bir formatla veri ambarlarında tutulmaktadır. Yapısal veri, veri tipine göre bir yapı içerisinde düzenlenmiş ve böylece tanımlanan veri, bir terim olarak kullanılır veya arama yapılabilir. Bunun aksine yapısal olmayan verinin tanımlanabilir bir yapısı bulunmamaktadır. Bilindiği üzere veri tabanlarında çok miktarda yapısal olmayan veri depolanmaktadır. Yapısal olmayan veri tipleri çoğunlukla; word ve text gibi metin dokümanları, resim dosyaları, pdf, Web üzerinde tutulan log dosyaları ve e-postalardır.

Bir veri madenciliği sisteminin kullandığı veri kümesinin farklı tipleri aşağıda gösterilmiştir:

1. Kayıt
   1) Veri matrisi (Data Matrix)
   2) Doküman verisi (Document Data)
   3) İşlem verisi (Transaction Data)
2.
   1) World Wide Web (WWW)
   2)Moleküler yapılar (Molecular Structures)
3. Sıralı
   1) Uzaysal veri (Spatial Data)
   2) Geçici veri (Temporal Data)
   3) Ardışık veri (Sequential Data)
   4) Genetik dizi verisi (Genetic Sequence Data)

Bu verilerden yazılım metodu ile bilgi çıkarımı önemli bir işlem konusudur. Burada, veri madenciliği, mevcut verilerden, aşikar olmayan ve önceden bilinmeyen fakat potansiyel olarak yararlı bilgilerin çıkartılması işlemi olarak değerlendirilir [10]. Bunun için veri özetleme, veri kümeleme, varyansların algılaması ve değişimlerin analiz edilmesi gibi birçok modele ihtiyaç vardır [30]. Çeşitli modeller oluşturmak ve bunlar içerisinden en iyi olanını seçmek için verilerin özelliklerinin daha iyi anlaşılması ve verilerin ön keşiflerinin yapılması gerekmektedir. Bu aşama genellikle veri hazırlama ile başlar. Bu süreçte; veri temizleme, veri birleştirme, veri dönüşümü, veri azaltma metotları kullanılarak, veri analiz için hazır duruma getirilir [3], [10].

Verilerin hazırlanma aşaması, aşağıdaki işlemlerin uygulanması ile tamamlanır:

1-Veri Temizleme: Eksik veriler tamamlanır, aykırı ve gereksiz verilerin silinmesi ile de gürültülü veriler veri tabanından temizlenir.

2-Veri Birleştirmesi: Bu aşamada, farklı veri tabanlarındaki veriler birleştirilerek tek bir ambarda depolanır ve uygulanacak işlemlere dahil olur.

3-Veri Dönüşümü: Düzeltme, birleştirme, genelleştirme ve normalleştirme gibi işlemler kullanılarak verinin, veri madenciliği metotları için uygun biçimlere dönüşümünü sağlar.

4-Veri İndirgeme (Seçme): Veri madenciliğinde en yüksek performansı elde edebilmek için büyük hacimli veri kümesinden daha küçük hacimli veri kümesinin oluşturulmasıdır.

bu işlemlerden sonra veriler, veri madenciliği için kullanılabilecek duruma getirilmiş olur.

Veri madenciliği algoritmaları açısından, metin veya Webdeki verilerin kalıplarını ortaya koymadan veya model oluşturmadan önce, verilerin yapısal hale dönüştürülmesinin gerektiğini belirtmeliyiz. Burada metin ve Web madenciliği yöntemleri, yapısal veriye ulaşmak için kullanılan araçlar olarak da ifade edilebilir [5], [30].

Kullanılan kayıtlar ve değişkenler, metin formatında olan verilerdir. Metinler bilgisayarın standart kullandığı veri formatında olmadığından dolayı bilgisayar bunları algılayamamaktadır. Ayrıca, her bir metnin dili ve içindeki anlam, onun kendi amacına yönelik olarak farklı şekillerde belirlenmektedir. Burada, yapısal olmayan bilgiden içerik çıkarmak için; anahtar kelimeler veya mantıksal aramalar, istatistiksel veya olasılıksal algoritmalar, sinir ağları ve kalıp keşfedici sistemler gibi dilbilimsel olmayan, geleneksel yöntemler kullanılırlar [8].

1.3. Veri Madenciliğinde Sınıflandırma Kavramı

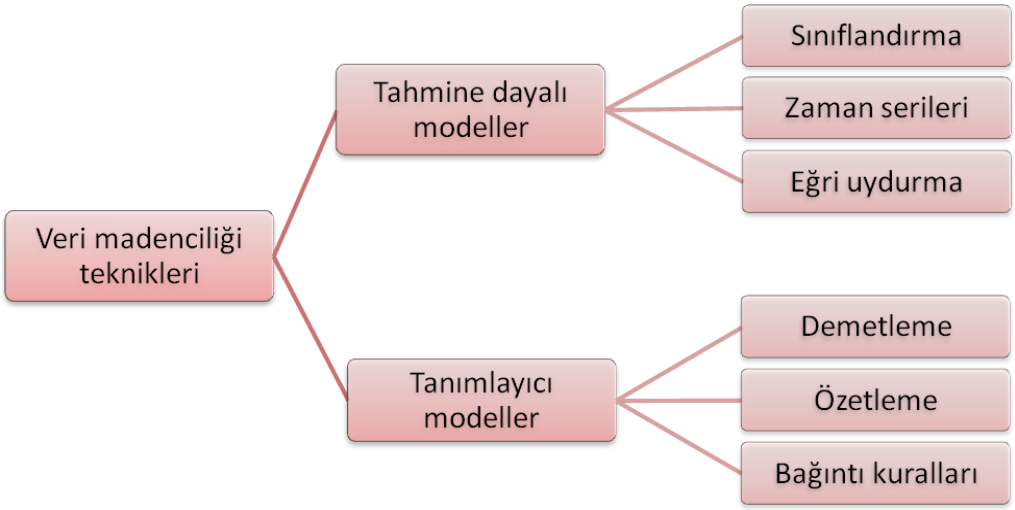
Sınıflandırma, yeni kayıtların doğru biçimde önceden biçimlendirilmiş sınıflara yerleştirilmesidir. Sınıflandırma modeli, sınıf özniteliği ile diğer niteliklerin değerlerinin

bütünlüğü işlemidir. Veri kümeleri, genel olarak eğitim ve test seti olarak ikiye ayrılır, eğitim seti ile model oluşturulurken test seti model doğrulama amacıyla kullanılır [16]. Sınıflandırma, bir grup veri içinde belli bir sınıf oluşturan nesnelerin benzer özelliklerine göre seçilerek gruplandırılması şeklinde tanımlanabilir. Otomatik sınıflandırmayla verilen bir nesne topluluğundaki benzer nesnelerin homojen sınıfları inşa edilir veya verilen nesnelerin özelliklerine göre matematiksel ve istatistiksel yöntemlerle önceden belirlenmiş sınıflarda toplanır. Sınıflandırma işlemi aşağıdaki basamaklardan oluşur;

- Model oluşturma: Model oluşturmak için kullanılan nesnelerin oluşturduğu veri kümesi, eğitim kümesi olarak adlandırılır.
- Model değerlendirme: Modelin başarımı (doğruluğu), doğru sınıflandırılmış test kümesi örnekleri kullanılarak belirlenir.
- Modeli kullanma: Model örneklerini sınıflandırmak ve onların nitelik değerlerini tahmin etmek için kullanılır.

1.4. Veri Madenciliğin Teknikleri

Veri madenciliği yöntemleri, çeşitli biçimlerde sınıflandırılabilir. Genel olarak veri madenciliği teknikleri tahmine dayalı (predictive) ve tanımlayıcı (descriptive) olarak iki grupta incelenir [14]. Şekil 3' de bu iki grup gösterilmiştir;



Şekil 3. Veri madenciliği modelleri

- Tanımlayıcı modellerin belirlenebilmesi için mevcut verilerde bulunan örüntüler kullanılır.
- Tahmin edici modeller, sonuçları bilinen verilerden hareket ederek bir model oluşturur ve oluşturulan bu model üzerinden sonuçları bilinmeyen veri setleri için sonuç değerleri tahmin edilir.

Başka bir biçimde veri madenciliği teknikleri, veri kümeleri üzerinde gözetimli (supervaized) ve gözetimsiz (unsupervaized) formda değerlendirilebilir. Değerlendirme metotlarına göre veri madenciliği teknikleri bölünür ise, sınıflandırma ve kümeleme yöntemleri olarak tespit edilir [21].

- Gözetimli (Supervised) : Bu yöntemde sınıfların sayısı ve hangi nesnenin hangi sınıfa ait olduğu bilinmektedir. Sınıflandırma yöntemleri bu metodu kullanarak çalışır.
- Gözetimsiz (Unsupervised): Bu yöntemde sınıf sayısı ve hangi nesnenin hangi sınıfa ait olduğu bilinmemektedir. Demetleme (clustering) yöntemleri bu metodu kullanarak çalışır.

Veri madenciliği sistemlerinde, veri sınıflandırma ve otomatik veri arama işlemleri için geliştirilmiş modeller vardır. Bu modeller aşağıdaki gruplarla tanımlanmıştır [6]:

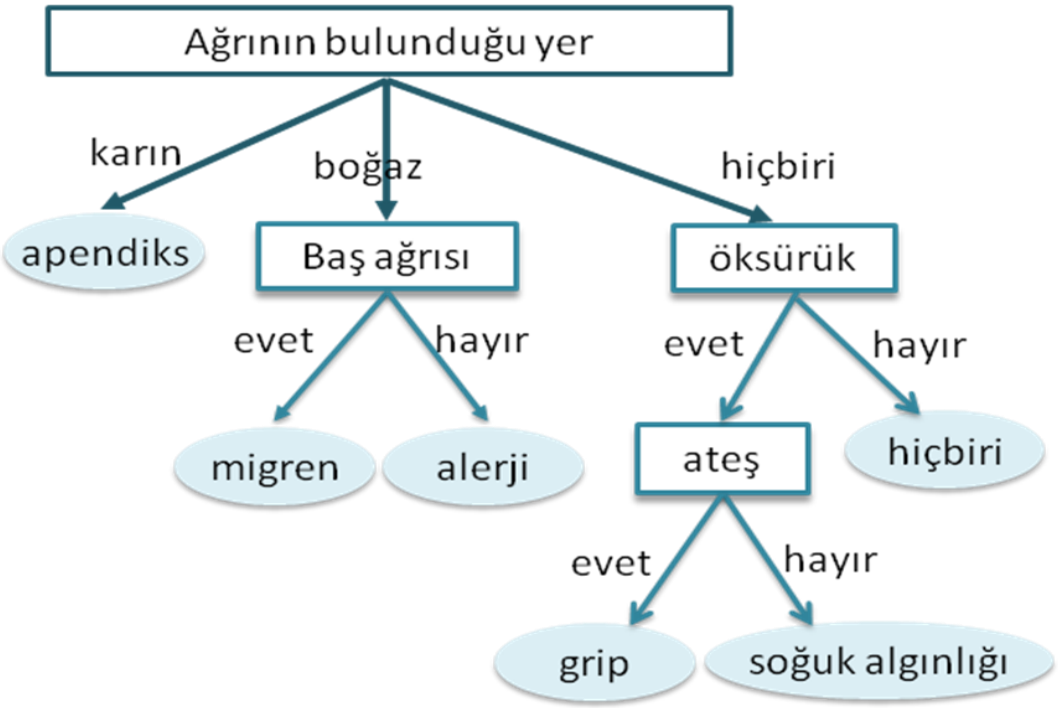
- Sınıflandırma (Classification)
- Kümeleme (Clustering)
- Birliktelik (Association)
- Dizi analizleri (Sequence Analysis)
- Sapma analizleri (Deviation Analysis)

1.4.1. Sınıflandırma

Sınıflandırma, eldeki yoğun veriyi analiz edip, nesnelerin özelliklerini kullanarak her nesneyi yine bu nesnelerin bir niteliği olan özel bir sınıfa atama işlemidir [2], [6]. Karar ağaçları (Decision Trees), Naïve Bayes, SVM (Support Vector Machines), bellek tabanlı sınıflandırma, yapay sinir ağları (Artificial Neural Networks), ve genetik algoritmalar en bilinen sınıflandırma teknikleridir.

1.4.2. Karar Ağaçları

Karar ağaçları çok tanımlanan, genel kullanımlı bir sınıflandırma metodudur. Ağaç şeklinde, yaprak düğümler ve sınama düğümlerinden oluşur ve veri üretildikten sonra ağacın kökünden yaprağına doğru Eğer-O kuralları (IF-THEN rules) uygulanır [10]. Kural oluşturma, veri madenciliği çalışmalarında sonuçların gerçeklenmesini sağlar. Bu kurallar sayesinde uygulama hususunda uzman bir kişiye gösterilecek olan sonucun tutarlı olup olmadığı sorgulanabilir. Daha sonra başka bir teknik uygulanacak olsa dahi, karar ağacı ile ön değerlendirme yapmak bize önemli değişkenler ve yaklaşık kurallar hakkında bilgi verir ve bizi yönlendirir. Şekil 4'de örnek olarak sanal nitelik ve değerlerden oluşturulan bir karar ağacını gösterilmektedir. Karar ağacının kök düğümünde Ağrının bulunduğu yer niteliği karın, boğaz ve hiçbiri değeri ile kıyaslanır, kıyaslamanın değerine göre ağaç farklı dallara ayrılıyor. Ulaşılan alt düğümlerde farklı nitelikler karşılaştırılarak, bir uç düğüme ulaşıncaya kadar aynı yöntem devam eder. Yaprak düğümlerde ise o düğüme ulaşan nesnelerin sınıfları yer alır.



Şekil4. Karar ağacı örneği

C4.5, J48 ve ID3 algoritmaları karar ağaçlarının kullandığı en bilinen uygulamalarıdır.

1.4.3. İstatistiksel Yöntemler

İstatistiksel sınıflandırma metodları Bayes teoreminden yararlanılır. Veri kümesinden her sınıfa ait ihtimal değeri niteliklere bağlı olarak hesaplanır. Oluşturulan bu ihtimallere göre bir nesnenin hangi sınıfa ait olduğu ihtimali olarak hesaplanabilir [6].

Veri madenciliğinde, istatistiksel sınıflandırma metotlarını kullanan algoritmalarının çoğu Bayes Teoremine dayalıdır. Çok sayıda istatistiksel sınıflandırma algoritması vardır. En yaygın olan ve sık kullanılan istatistiksel veri madenciliği algoritmaları; Naive Bayes Algoritması ve Bayes Ağlarıdır. Bu teknikler istatistik literatürde çok boyutlu analiz (multivariate analysis) başlığı altında toplanır ve genelde verinin parametrik bir modelden (çoğunlukla çok boyutlu bir Gauss dağılımından) ortaya çıktığını farz ederler. Bu varsayım adı altında uzun yıllardır sınıflandırma (classification; discriminant analysis), regresyon, öbekleme (clustering), boyut azaltma (dimensionality reduction), hipotez testi, varyans analizi, bağıntı kurma (association; dependency) gibi istatistikte teknikler kullanılmaktadır [28].

1.4.4. Bellek Tabanlı Yöntemler

Normal hayata ilk defa gördüğümüz bir nesnenin ne olduğunu anlamak için hafızamızda yer alan eski nesnelerle karşılaştırırız ve bu yeni nesne en çok herhangi nesneyi anlatıyorsa bu nesneyi de onunla aynı sınıfa atarız. Bellek tabanlı metotlar, yeni bir nesnenin sınıfını belirlemek için bu nesnenin öznitelikleri ile eldeki nesnelerin öznitelikleri arasındaki benzerlikler veya farklılıkları göre yeni nesneyi en çok benzediği sınıfına atar. Nesnelerin arasındaki benzerlikler, öznitelikler arasında uzaklık ölçümüne göre matematiksel olarak değerlendirilir. Bu metodun en iyi örneği, *k* - en yakın komşu algoritması (*k*-nearest neighbor) olarak belirtilmiştir [10].

1.4.5. Yapay Sinir Ağları

1980 yıllarından sonra yaygınlaşan yapay sinir ağlarında (artificial neural networks) asıl fonksiyon birbirine bağlı basit işlemci ünitelerinden oluşan bir ağ üzerine yayılmıştır [14].Yapay sinir ağlarında kullanılan öğrenme algoritmaları veri ile üniteler arasındaki bağlantı ağırlıklarını ölçer. Yapay Sinir Ağları istatistiksel yöntemler gibi veriyi parametrik bir model olarak varsaymaz, yani daha geniş uygulama alanına sahip ve bellek tabanlı modeller kadar yüksek işlem ve bellek gerektirmez.

1.4.6. Kümeleme

Kümeleme, birbirinden çok farklı özelliklere sahip olan kümelerin tespit edilmesini sağlayan bir yöntemdir. Veri kümesi birbirine benzeyen nesnelerden oluşan kümelere bölünür. Aynı kümedeki veriler birbirine daha çok benzerliklere sahip olup, farklı kümelerdeki nesneler ise birbirine daha az benzerler. Bazı çalışmalarda kümeleme işlemleri, sınıflandırma yöntemin önişlemi olarak da uygulanmaktadır. Kümeler içinde yer alan elemanlar, birbirlerine benzer özellikler göstermektedirler ve veriler herhangi bir sınıf içerisinde yer almaz. Bu uygulamaya örnek olarak, alış veriş merkezlerinde, farklı müşteri gruplarının bulunması ve bu grupların alışverişle ilgili desenlerinin keşfedilmesi verilebilir. Kümeleme yöntemlerini aşağıdaki verilmiştir [10], [12] .

- Bölme yöntemleri (Partitioning methods)
- Hiyerarşik yöntemler (Hierarchical methods)
- Yoğunluk tabanlı yöntemler (Density-based methods)
- Grid tabanlı yöntemler (Grid-based methods)
- Model tabanlı yöntemler (Model-based methods) [14].

Kümeleme yöntemi, aynı karakteristik özelliklere sahip olan nesnelerin bir araya toplanması sürecidir. Bu yöntem Web madenciliği için, genelde Kullanıcı Grupları (User Clusters) ve Sayfa Grupları (Page Clusters) olarak iki küme yaklaşımı kullanılmaktadır.

1.4.7. İlişkilendirme Kuralları

İlişkilendirme kuralları (Association Rules) veya birliktelik kuralları analizi aynı zamanda pazar sepeti analizi olarak da adlandırılabilir. Bu teknikle eş zamanlı olarak meydana gelen olaylar incelenir [21]. Örnek olarak, bir müşterinin bütün alışverişlerde satın almış olduğu ürünlerin arasındaki ilişkileri tespit edilerek müşterinin satın alma alışkanlıkları analiz edilebilir. Müşterilerin hangi ürünleri bir arada aldıkları ile ilgili bilgilerin ortaya çıkartılmasıyla market yöneticileri, bu bilgiler sayesinde daha etkin satış stratejileri geliştirebilirler. Büyük veri tabanlarında birliktelik kuralları tespit edilirken şu iki işlem yapılmalıdır:

a. Sık tekrarlanan öğelerin bulunması,

b. Sık tekrarlanan öğelerden güçlü birliktelik kurallarının oluşturulması [13], [26].

Genellikle alışveriş işlemlerinde kullanıldığından dolayı İlişkilendirme Kuralları aynı zamanda alış veriş sepeti analizi olarak da bilinmektedir. Bu yöntemdeki amacı, bir küme içerisindeki nesnelerin birbirleri ile olan ilişkilerini belirlemektir. Veri Madenciliğinin bu yönteminin yaygın olarak alışveriş sistemlerinde kullanıldığı görülse de bu yöntem başka uygulamalarda da kullanılmaktadır. İlişkilendirme kuralı yöntemiyle A ürünü ile B ürününün veya C ürününün alınması arasında bir bağlantı olup olmadığının tespit edilmesi ve eğer bağlantı varsa bu bağlantılar arasındaki kuvvet veya önem derecesinin (confidence or strength) tespiti sağlanır. Bu analizin amacı, A ürününü alan kişilerin B veya C ürünleri alımları arasında kuvvetli bir ilişkinin bulunmasıyla sistemde bir takım değişiklikler gerçekleştirmektir. Örneğin, alışveriş sistemlerinde çeşitli promosyonların düzenlenmesi, ürün raflarının elde edilen sonuçlar doğrultusunda düzenlenmesi yapılmaktadır. Bu işlem, bir Web sitesi içerisindeki sayfaların şekillendirilmesinde de kullanılır [25], [28].

1.4.8. Dizi Analizleri

Bir dizi, farklı değerdeki serilerinden oluşur ve dizi analizleri (Sequential Patern) de bu farklı serilerde örüntüler bulmak için kullanılan yöntemlerdendir. Bir DNA dizisinin A,

G, C ve T gibi 4 farklı durumun farklı dizilmesiyle meydana gelen serilerin birleşimi olması buna iyi örnektir [14].

Dizi analizleri ve birliktelik kuralları analizleri arasında, belirli durumların kümeleri üzerinden işlem yapmaları yönünde bir benzerlik vardır denilebilir. Ancak, dizi analizleri, durumlar arası geçişleri analiz ederken birliktelik kuralları analizleri eş zamanlı ve birbirinden bağımsız oluşan durumları inceler [27]. Sıralı doku (pattern) yöntemiyle ilgili kullanıcı oturumları arasında doku kurulmaya çalışılır. Bu yöntemde, belirli zaman aralıklarında oturumlar ele alınır ve bunlar arasında karşılaştırmalar yapılır. Bunun için sıralı doku yönteminde, eğilim analizi, değişen nokta bulma veya benzerlik analizleri gibi bazı geçici analiz türleri kullanır. Bu yöntemin kullanılması ve sonuçları, gelecekteki eğilimi tahmin etmek isteyen Web pazarlamacıları için oldukça önemlidir. Bu sayede hazırlanacak ilanlar belirli kullanıcı gruplarına göre düzenlenir.

1.4.9. Sapma Analizleri

Milyonlarca işlemde normal olmayan durumların tespit edilip tanımlanması oldukça zor bir işlemdir. Diğerlerinden farklı seyir gösteren bu anormal durumları ortaya çıkarmak için de sapma analizlerine başvurulur. Bu yöntem daha çok kredi kartı yolsuzluklarının ortaya çıkarılması sürecinde kullanılır. Bunun yanı sıra, bir ağın gereksiz meşgul edilip edilmediğini denetlerken ve üretim hatalarını incelerken de kullanılabilir. Ancak bu yöntem sadece görselleştirme veya istatistiksel tekniklerle uygulanabilir. Analiz işlemi için kullanılabilecek bir başka yöntem de doğrusal regresyon yöntemidir. Bu yöntemin en çok bilinen uygulaması istisna sapmasıdır. İstisna sapması, kredi kartı yolsuzluklarının tespiti için yaygın olarak kullanılan yöntemlerden biridir [17]. Sapma analizi üzerindeki çalışmalar devam etmekle beraber bunun için henüz standart bir teknik geliştirilememiştir [27].

1.5. Sınıflandırma Algoritmaları

1.5.1. Naive Bayes Sınıflandırıcı

Naive Bayes algoritması sınıflandırma yönteminin basit bir olasılık algoritmasıdır ve Bayes kuralına göre güçlü bir bağımsızlık varsayımlarına dayalı uygulanır. İstatistiksel yöntemler vasıtasıyla sınıflandırma yapan bu yöntem, hızlı ve kolay bir şekilde uygulanabildiği ve herhangi bir karmaşık parametre içermediğinden dolayı oldukça önemlidir. Naive Bayes algoritmasının uygulanmasında en önemli kural niteliklerin birbirinden bağımsız olduğudur. Niteliklerin birbirini etkilemesi durumunda olasılığın hesaplanması zorlaşır. Bu durumda sadece öznitelikler arasında bağımsızlık olduğu farz edilerek Bayes modeli uygulanabilir. Bu metodun esası, her veri için bir olasılık dağılımı varlığı prensibine dayanmaktadır, ki yeni bir veri ortaya çıkması ile onun olasılık dağılımı hakkında optimal kararlar benimsenebilir.

Bayes teorisinin temel taşını Bayes öğrenimi oluşturmaktadır. Bu teori başlangıç olasılıklara dayalı ikinci olasılıkların hesaplamasını mümkün kılar. Diyelim ki H bir hipotez uzayı ve D eğitim örnekleri olarak mevcut olsun. Burada, Bayes kuralı Eşitlik 1'de olarak ifade edilir.

$$p(H|D) = \frac{p(D|H) * p(H)}{p(D)} \quad (1)$$

Bayes yönetiminin ana düşüncesi, bir hipotez veya bir olayın (H) sonuçlarını, tespit edilen bazı kanıtlara (D) göre tahmin edilebilmektir.

1. Bir önsel olasılığı H veya P(D|H): kanıtlar tespit edilmeden önceki bir olayın olasılığıdır.
2. Bir sonsal olasılığı H veya P(H|D) den: kanıtlar görüldükten sonra bir olayın olasılığıdır.

   P(D) ve P(H) sırasyila eğitim örneklerinin ve bir hipotez uzayın ihtimaleridir.

Görüldüğü gibi P(D) miktarı artması ile P(H|D) miktarı azalır. Çünkü, ne kadar H den bağımsız varsayan D in görülme olasılığı daha fazla olursa, H destekleyen D deki kanıtların az olduğu anlamına gelir.

Naive Bayes modeli Bayes teorisinin çok pratik bir uygulaması olarak tanımlanmaktadır. Bu model, metin sınıflandırma ve medikal teşhisler gibi uygulamalarda, sinir ağları ve karar ağaçları ile karşılaştırıldığında iyi performans verebilmektedir. Bunun yüzünden, metinsel dokümanlarının sınıflandırılmasında genellikle bu yöntemden yararlanılır. Ayrıca, Naive Bayes modeli birçok farlı uygulamalarda etkin olarak kullanılmaktadır. Herhangi bir uygulama, aşağıdaki dört kuralı sağlıyorsa Naive Bayes yöntemini kullanabilir [3]:

1. Özelliklerin bağlaçları tarafından X (x_1, x_2, ..., x_n) örneğin tanımlanabilmesi. X Örneklerin kümesini temsil eder.
2. Özellikler arası bir koşul olarak, özellikler birbirinden bağımsız olmalıdır.
3. F(x) objektif fonksiyonu, sınırlı sayıda olan V kümesi içindeki her değeri alabilmelidir.
4. Oldukça büyük eğitim örnekler seti mevcut olmalıdır.

Naive Bayes yöntemi Bayes teorisine göre aşağıdaki şekilde hesaplanabilir.

F(x) objektif fonksiyonunu, f: X→V olarak düşünürüz ve ondaki her x örneği, (a_1, a_2, ..., a_n) özellikler tarafından belirlenir. Bayes yaklaşımında problem çözmek için f(x) deki en büyük olasılığı V_{map} hesaplanır [4], [32].

$$v_{map} = arg \max_{v_j \in V} P(v_j | a_1, \dots, a_n) \tag{2}$$

Denklem (2) Bayes eşitliği kullanılarak, aşağıdaki gibi yazılır.

$$v_{map} = arg \max_{v_j \in V} \frac{P(a_1, \dots, a_n | v_j) P(v_j)}{P(a_1, \dots, a_n)} \tag{3}$$

$$v_{map} = arg \max_{v_j \in V} P(a_1, \dots, a_n | v_j) P(v_j) \tag{4}$$

Yukarıdaki denklemde (4), $P(v_j)$ değerinin, kaç defa v_j eğitim örnekleri kümesinde var olduğunu sayarak, hesaplanır. Diğer yandan, $P(a_1, a_2, \ldots, a_n \mid v_j)$ hesaplanması çok pratik değildir. Ancak, $P(a_1, a_2, \ldots, a_n \mid v_j)$ çok büyük eğitim veri seti mevcut ise hesaplanabilir.

Eşitlik (4), özelliklerin birbiriyle bağımsızlıkları koşulunu dikkate aldıktan sonra, belirtilen f(x) için, bağlaç $(a_1, a_2, \ldots, a_n)$ görme olasılığı, bireysel özellikler olasılıklarının çarpımlarına eşittir. Bu durumda, eşitlik (4) aşağıdaki şekilde hesaplanabilir:

$$v_{NB} = arg \max_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(a_i|v_j) \quad (5)$$

Naive Bayes modelinin metin alanında daha iyi uygulanabilmesinin sebeplerinden biri, kanıtların metinde yer alan "kelime" veya "sözcükler" olmasıdır. Genelde sözlüklerin boyutu binlerce farklı aralıkta yer alır. Kanıtların veya sözcüklerin boyutunun fazla olması, metin sınıflama probleminde Naive Bayes modelinin sağlıklı çalışmasını sağlayan bir etkendir. Bu yöntem, metin sınıflandırılmasında, terimlerin belge içerisindeki dağıtımını hesaplayarak yeni gelen belgeler için sınıf tahminini yapabilir [10]. Bu tahmini yapabilmesi için aşağıdaki kuralların uygulanması gerekir:

1. Naive Bayes modelini inşa etmek
2. Metin dokümanlarının sınıflandırılması

Metin dokümanlarının sınıflandırılması için örnek olarak Web'deki metin sayfalarının hangi konuda üzerinde olduğunun belirlenmesi verilebilir. Böyle bir uygulama için Bayes sınıflandırma yöntemi, özelliklerin birbirlerinden bağımsız olmaması durumunda bile çok etkili olarak çalışabilir.

Bir metinsel verinin öznitelik değerleri olarak gösterilmesi için iki yol denenmiştir:

1. Metin içindeki her kelime pozisyonu bir özellik olarak kabul edilir. Örneğin, 100 kelimeden oluşan bir metin aynı zamanda 100 özellik içerir. Bu yöntemde, mevcut olan her metinsel doküman kelimelerden oluşan bir vektöre dönüşür ve her kelimenin pozisyonunun karşılığında bir özellik atfedilir ki bu öznitelik değeri kelime ile aynı olacaktır. doc = $(a_1=w_1, a_2=w_2, \ldots, a_n=w_n)$.

2. Sözlükte mevcut olan her kelime (örnek: yaklaşık 50000), bir özellik olarak değerlendirilir ve metinde tekrar sayısı sayılır.

Metin sınıflandırma Bayes kuralına göre aşağıdaki şekilde hesaplanır:

$$P(C_i|B) = \frac{P(B|C_i)P(C_i)}{P(D)} \tag{6}$$

B = (t_1 .... t_n) terim vektörü ile temsil edilen bir belge için, $P(B|C_i)$ ihtimali formül (7) ile hesaplanabilir [15].

$$P(B\ |C_i) = \prod_{j=1}^{j=n} P\ (t_j\ |S_i\ ) \tag{7}$$

Elde edilen bu bilgiyi kullanmak için, bir belgenin dahil olabileceği sınıfı bulmakta ve daha çok tercih edilecek bu sınıfa daha çok şans tanımak uygun bir yöntem olabilir [15].

$$P(S_i) = \frac{sınıftaki\ eğitim\ belgeleri\ sayısı}{toplam\ sınıf\ \ sayısı} \tag{8}$$

Sonunda M adet sınıf varsa, bir sınıf seçme işlemi formül (9) ile hesaplanabilir.

$$argmax_{s_i}\ [P(S_i|B)] = \ argmax_{s_i}[P(B|S_i)P(S_i)] \tag{9}$$

Naive Bayes algoritması, belirli bir sınıf için terim ihtimallerini hesaplama yöntemini, çok terimli (multinominal) ve çok değişkenli (multivariate) olmak üzere iki farklı şekilde uygulanır.

Çok terimli yöntemde terimlerin ne kadar tekrar ettiği dikkate alınır. Buna karşın çok değişkenli metotta sadece terimlerin var olup olmadıklarına bakılır. Burada, Naive Bayes algoritması bit ağırlıklandırma ve frekans ağırlıklandırma yöntemi olarak işlem yapar.

1.5.1.1. Naive Bayes Bit Ağırlıklandırma Yöntemi

Aşağıda belirtilen 10 ve 11 formülleri ile d vektörünün c_j kategorinde olma ihtimali hesaplanır.

$$P(d|c_j) = \prod_{t=1}^{|V|} P(w_t|c_j)^{x_i}(1 - p(w_t|c_j))^{(1-x_t)} \tag{10}$$

$$p(w_t|c_j) = \frac{1 + B_{jt}}{2 + |c_j|} \tag{11}$$

Burada $|C_j|$, $|V|$, B_{jt}, X_t anlamları, sırasıyla c_j sınıfında bulunan eğitim dokümanı sayısı, sözlükteki kelime sayısı, c_j kategorisinde bulunan ve w_t kelimesini içeren eğitim dokümanın sayısı ve kelimenin ağırlığı (1veya 0), anlamlarına gelirler. Formül 12'e göre M(C) değerinde en büyük olan sınıfa aittir [26], [32].

$$\mathrm{M(C)} = \mathrm{P(word\ 1|C)^{n1}P(word\ 2|C)^{n2}} \ldots \mathrm{P(word\ v|C)^{nv}P(C)} \tag{12}$$

1.5.1.2. Naive Bayes Frekans Ağırlandırma Yöntemi

Aşağıdaki denklemler ile Naïve Bayes algoritmanın multiominal modeli oluşturulur:

Eşitlik 13 ve 14 da d kategori sayısını, P(|d|) kategori olasılığı ve X_t kelimenin frekansını temsil eder.

$$p(d|c_j) = p(|d|)|d|! \prod_{t=1}^{|v|} \frac{p(w_t|c_j)^{xt}}{x_t!} \quad (13)$$

$$p(w_t|c_j) = \frac{1 + N_{jt}}{|V| + N_j} \quad (14)$$

Bu formülleri ifade eden N_{jt}, N_j ve $|V|$ değerleri sırasıyla, j sınıfındaki dokümanlar içinde t kelimesinin tekrarlanma sıklığı, j sınıfındaki toplam kelime sayısıdır. Naïve Bayes bit ağırlıklandırma yöntemi M(C) bağlı olarak belirlenir ve örnek M(C) değeri en büyük olan sınıfa atanır.

Çok terimli (multinominal) Naive Bayes algoritması çok değişkenli multivariate Naive Bayes algoritmasına göre daha iyi sonuçlar verdiği görülmüştür. Burada dikkate alınması gereken mesele, her bir kelimenin tekrarlanma sayısının diğer kelimelerin tekrarlanma sayılarından bağımsız olmasıdır [10], [14], [26].

1.6. Veri Madenciliği Alanları

Veri madenciliğinde çok kullanılan süreçlerden biri de metin ve Web madenciliği işlemleridir. Bunlar veri madenciliğinde yapısal veriyi elde etmek için kullanılan yollar olarak da ifade edilmişlerdir. Son birkaç yıldır metin ve Web madenciliği büyük oranda birbirine bağlı olarak bir arada çalışılan alanlardır.

Metin madenciliği, çok büyük belgelerin analiz edilmesi ve metin tabanlı verinin içerisindeki gizli kalıpların ortaya çıkarılmasıdır. Web madenciliği ise, Web içerikleri, sayfa yapıları ve Web bağlantı istatistiklerinin de içerisinde yer aldığı Web ile ilişkili olan Verilerin analizini kapsamaktadır [16], [29].

1.6.1. Web Madenciliği

Son zamanlarda birçok çalışmanın internet üzerinde düzenlenmesi sebebiyle çok büyük oranlarda veri dağılımı ortaya çıkmakta ve bunlar www (World Wide Web) ortamda kullanım halindedir. Web madenciliği hızlı büyüyen bir araştırma sahasıdır. Web ortamında veriler çok farklı standart ve biçimlerde yer almaktadır. Veri dağıtımları aşağıda gösterildiği gibi farklı biçimlerde ve farklı tiplerde olabilir [3].

- Web sayfaları
- Kullanıcı kayıt bilgileri
- Site yapısı ve içeriği
- Log erişim (Access Log) dosyaları
- Oturum ve hareket bilgileri

Web madenciliği, yukarıda tanımlanmış çeşitli yapılarda olan Web sayfaların dokümanlarını ve kayıt bilgilerini inceleyip bunlardaki kalıpları ortaya çıkarmak için veri madenciliği tekniklerinin kullanılması olarak tanımlanabilir [29].

Veri madenciliği tekniklerinin world wide Web verileri üzerinde uygulanması, Web madenciliği olarak ifade edilir. Web madenciliği aşağıdaki verilen üç farklı alt başlıkta incelenebilir:

1. Web içerik madenciliği
2. Web yapı madenciliği
3. Web kullanım madenciliği

1.6.1.1. Web İçerik Madenciliği

Web içerik madenciliği, veri ve metin madenciliğine bağlı olmakla beraber bazı yönlerden onlardan farklılık arz eder. Web içerik madenciliği, veri madenciliğiyle ilgili olduğu için veri madenciğinin birçok tekniğini uygulamaktadır ve ayrıca Web sayfalarının içeriğini daha çok metinler oluşturması bakımından metin madenciliği ile de bağlantılıdır [29]. Ancak bunlarla birlikte, bu kavramın veri madenciliğinden farklı olmasının sebebi ise

Web verilerinin çoğunlukla yarı-yapılandırılmış veya yapısız olmalarıdır. Aslında veri madenciliği sadece yapısal veri ile ilgilenir. Buna ek olarak Web'in doğası yarı-yapılandırılmış metinler iken metin madenciliği sadece yapılandırılmamış metinler üzerine odaklanmaktadır. Böylece Web içerik madenciliği, yapay zekâ, akıllı yazılım programları ve bilgi tarama tekniklerini kullanarak Web kaynaklarının içeriklerinden yararlı bilgiyi elde etmeye çalışmaktadır.

Web içinde farklı yapılarda olan veriler (metin, görsel, link, resim ve benzeri) Web içerik madenciliği için yapılacak uygulamaları zorlaştırır. Web sitelerinin belgelerindeki linkleri ve hyperlinkleri bularak, sayfanın ve Web sitesinin yapısal raporunu çıkarmaya çalışır.

Web içerik madenciliğinde, eldeki çalışmanın amacına göre üç farklı rapor oluşturabilir [28]:

- Web sayfasının hyperlinklere bağlı olarak sınıflandırılması
- Belirli bir alan adının Web sitesindeki yapısal hiyerarşisi ve hyperlink ağının raporu
- Web sitesi yapısını gösteren rapor

Web içerik madenciliğinde elde edilen sonuçlar kullanıcıların bilgi arayışlarında yararlanabilecekleri görsel sunumlara çevrilir.

1.6.1.2. Web Yapı Madenciliği

Web yapı madenciliği, Graph Teorisi kullanılarak bir Web sitesinin düğümünün çözülmesi ve bağlantı yapısının analiz edilmesi için kullanılan işlemdir. Başka bir ifadeyle, Web yapı madenciliği tasarımı Web sayfaları arasındaki linkleri takip ederek bilgi üretmektir. Web yapı madenciliği, yapısal verilere göre iki şekilde yapılır [5]:

- Webdeki hyperlinklerin modelinin çıkarılması: Hyperlink, bir Web sayfasını farklı bir lokasyona yönlendiren yapısal elemandır.
- Belge yapısının madenciliği: Web sayfası dokümanlarındaki HTML veya XML etiketlerinin açıklanması için ağaç ve benzeri yapıların kullanılması ve analizidir.

1.6.1.3. Web Kullanım Madenciliği

Web kullanım madenciliği, log dosyalarından veya kullanıcıların geçmiş hareketlerinden faydalı bilgiler ayıklama işlemidir. Web kullanım madenciliği, kullanıcıların internet üzerinde aradıklarının ne olduğunu bulma sürecidir. İstemcilerden gelen her istek, bir kayıt olarak, metin tabanlı log dosyalarına ilave edilir. Bu log dosyalarının kayıt desenlerindeki veriler, kullanıcı hakkında, ayrıntılı bilgiler içerir. Log dosyasındaki kayıt formatı, verilen servis çeşidine ve kullanılan işletim sistemine göre farklılıklar gösterir. Bu log dosyalarından bazıları şunlardır: access log (erişim), mail log, error log, referrer log, ftp log [5], [7] Bunların haricinde, sunucu üzerindeki verilen farklı servislerde isteye bağlı olarak log dosyaları bulundurulmaktadır. Özellikle Web sunucularının (Apache, Microsoft II S) access log dosyaları, içerdikleri veriler nedeniyle Web madenciliğinde ciddi bir veri kaynağı vazifesi görmektedirler [24].

İnternet kullanımının günümüzde ciddi bir şekilde artış göstermesiyle Web madenciliği hakkındaki yapılan çalışmalar, her geçen gün artmaktadır.

1.6.2. Metin Madenciliği

Metin madenciliği, doğal dilde yazılan metinlerden anlamlı ve nitelikli bilgilere ulaşmaya çalışan yeni bir çalışma alanıdır. Metinlerin sınıflandırılması, bu alanda yapılan önemli çalışma alanlarından biridir. Metin sınıflandırma, doğal dilde yazılı halde bulunan belgelerin içeriği ile ilgili olarak önceden belirlenmiş sınıflara dahil edilmesi işlemine verilen isimdir [10]. Başka bir açıdan, metin sınıflandırma, belgelerin sahip olduğu özelliklere göre, önceden belirlenmiş kategorilerden hangisine dahil olacağının tespit edilmesidir. Metin sınıflandırmanın, bilgi alma (information retrieval) veya bilgi çıkarma (information extraction), doküman indeksleme veya filtreleme, otomatik olarak meta-data elde etme ve Web sayfalarının hiyerarşik olarak düzenlemesi gibi pek çok alanda önemli rollere sahiptir [2], [18]. Herhangi bir kaynaktan alınan haberlerin konularına göre sınıflandırılması, metin sınıflandırma işlemi için güzel bir örnek olarak ifade edilebilir. Ancak, doğal dilerdeki metin yazılımları, bilgisayar için uygun bir yapısal veri olmadığından bununla ilgili her türlü işlemin bilgisayarda yapılabilmesi için harfler ve

kelimelerin matematiksel olarak işlenmeye biçime çevrilmesi gerekir. Diğer bir ifadeyle metin veya Web verileriyle herhangi bir işlem yapılmadan veya model oluşturulmadan önce, veri madenciliğinde kullanılabilmeleri için verilerin yapısallaştırılması gerekir. Burada, veriler, farklı şekillerde bulunabilir. Bazıları otomatik veri analiziyle çözümlenmeye uygun iken bazılarının analizi oldukça zordur. Klasik veri analiz yöntemleri verinin değişken olduğu ve kayıt bazlı düzenlendiği ve olasılığı ile işlem yapmaktadır. Buradaki sorun, verinin metin formatında yani kayıtların ve değişkenlerin olmadığı bir yapıda olması durumunda, yapılması gerekenin ne olduğudur.

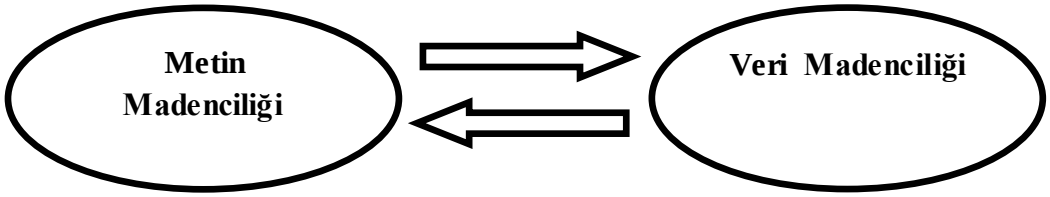
Metin yazımında standart kurallar olmadığından dolayı bilgisayar bunları algılayamamaktadır. Amacına yönelik olarak her bir metnin dili ve içerdiği anlam, çeşitlilik göstermektedir. Yapısal olmayan bilgiden içerik çıkarmak için anahtar kelimeler, mantıksal aramalar, istatistiksel veya olasılıksal algoritmalar, yapay sinir ağları ve kalıp keşfedici sistemler gibi dilbilimsel olmayan geleneksel yöntemler kullanılmaktadır. Bu yöntemlerin temeli, hem sorgudaki hem de metindeki kelimelerin karakterlerini karşılaştırmaya dayanır. Bu sebeple orijinal metnin içeriğinden doğrudan açıklayıcı sonuçlara ulaşılamaz. Bir doğal dilin anlamsal temeli, dilbilimsel esaslara dayanır ve bu genellikle Natural Language Processing (NLP) olarak isimlendirilir. Bir NLP sistemi, karışık ifadelerin bulunduğu yapıları (örneğin; duştan akan soğuk su ile içilen soğuk su arasındaki fark gibi) mantıksal olarak terimleri sınıflayarak; ürünler, organizasyonlar veya kişiler gibi gruplara dönüştürmektedir. Doğal dil metinlerinden anlamlı ve nitelikli bilgi elde edilmesini sağlayan metin madenciliği, iki aşamada gerçekleştirilir [12], [13] :

- Anahtar içerik/ifadeler metinden elde edilir,
- Elde edilen içerik/ifadeler, büyük oranda ilişkili olduğu gruplara ayrılır.

Metin madenciliğinde işlemleri iki ana grupta toplanabilir:

- Metnin anlaşılması/özetlenmesi: Metin madenciliğinin amaçlarından birisi de metinden anlamlı ve nitelikli bilginin çıkarılmasını sağlamaktır. Böylece metnin içinde bulunan anahtar içerik belli olacaktır.
- Metin ile modelleme: içerdiği anahtarlar ile tahmin edildiği bir modelin geliştirilmesi aşamasıdır. Burada, metinden elde edilen içerik girdi değişkeni olarak kullanılır ve diğer bilgiler ile beraber öngörüsel model geliştirilir.

Veri madenciliği, girdi olarak sadece yapısal veriyi kullandığı için veri madenciliğinin çözümleri ve algoritmaları kullanılarak metin verisinden kalıplar bulunması, modeller kurulmadan önce de metinden elde edilecek bilginin yapısal hale dönüştürülmesi gerekmektedir. Metin madenciliği sayesinde, kategorilerin oluşturulmasıyla yapısal olmayan veri yapısal hale getirilmektedir [11], [13]. Şekil 5'de metin madenciliği ve veri madenciliği arasındaki ilişki gösterilmiştir.



Şekil **5.** Süreçler Arasındaki ilişki

Şekil 5'de görüldüğü gibi, metin ve veri madenciliği arasında etkileşimli bir ilişki vardır. Metin madenciliği sonucunda elde edilen yapısal veriler, veri madenciliği modellerinde ve elde edilen sonuçlar ise daha sonra metnin yapısının incelenmesinde kullanılmaktadır.

Metin madenciliğinin uygulama alanlarından bazıları şunlardır [2], [8] :

- Müşteri ilişkileri yönetimi (Customer Relationship Management, CRM): Bütün müşterilerin E-mail, işlem, çağrı merkezi ve anket gibi erişim noktalarındaki metin bilgilerinden nitelikli bilgi ortaya çıkarılır. Bu nitelikli bilgiler de müşterinin terk etme ve çapraz satışlarını tahmin etmek için kullanılır.
- Sahtekârlık (Fraud) keşfi: Sağlık, sigorta ve hükümet tarafından toplanan büyük miktardaki metin verilerindeki kalıplar ve sıra dışılıklar aranarak bunlardaki hileler belirlenir.
- Bilimsel ve medikal incelemeler: Makale başlıkları, yayınlanmış araştırma sonuçları, hasta raporları ve diğer yayınlardaki metin materyallerinden tespit yapılır.
- Güvenlik/istihbarat: Organizasyonları, terörist tehlikelerini, suçlu (criminal) davranışları, bireyler arasındaki kalıpları ve bağlantıları tahmin etmek ve engelleyebilmek için çok miktarda metin içerisinde arama yapılır.

- Pazar araştırması: Yayınlanmış belgeler, basın bültenleri ve Web sayfaları pazar etkisinin ölçülmesi için aranır ve izlenir. Aynı zamanda, metin madenciliği niceleyici yöntemler ile açık uçlu anket soruları ve mülakatların değerlendirilmesinde de kullanılmaktadır [6].

Bir kaynaktan alınan bilgilerin konularına göre sınıflandırılmaları metin sınıflandırmanın önemli uygulama alanlarından biridir. Ancak, doğal dilerdeki metinler bilgisayar için doğru yapısal bir veri değildir ve bilgisayarla veri madenciliğinde işlenebilmesi için metini oluşturan harflerin ve kelimelerin matematiksel bir şekle çevrilmesi gerekir. Böylece metinsel verile veri madenciliğinde işlenebilecek yapısal biçimi dönüştürülürler.

1.6.2.1. Metin Sınıflandırma

Bir sınıfın oluşabilmesi için her belgenin belli ortak özelliklere sahip olması gerekir. Ortak özelliklere sahip olan belgelerin hangi özellikleriyle bu sınıfa dahil olacağını tespit eden algoritma, sınıflama algoritması olarak tanımlanmaktadır. Sınıflama algoritması, denetlenen öğrenme kategorisine dahil olan bir öğrenme yapısıdır. Denetimli öğrenme ise, hem girdi hem de çıktıyı içeren öğrenme ve test verilerinin kullanmasıdır. Sınıflama ile amaçlanan, bir belgenin önceden tespit edilmiş bir sınıfa dahil edilmesidir [21], [31].

Bir belgenin önceden tespit edilmiş bir gruba dahil edilebilmesi için de sınıflama algoritması ile öğrenme verileri kullanılarak hangi sınıfların var olduğu ve bu sınıflara girmek için bir belgenin hangi özelliklere sahip olması gerektiği otomatik olarak belirlenmelidir. Bununla beraber test verileriyle de bu öğrenmenin testi yapılarak ortaya çıkan kurallar optimize edilmelidir.

Metin sınıflandırma, mevcut sınıflardan birine ait olduğu bilinen bir dokümanın, hangi sınıfa ait olduğunun tespit edilmesi işlemidir. Günlük hayatta bir gazete ya da bir kitap okunduğunda, bu metinlerde gecen olaylar daha önce bilinen birtakım olaylarla ilişkilendirilir. Burada, bilgilerin kendi aralarında nasıl bağlandığı ve her sınıfın içinde yer aldığı konulara bakılarak bir konunun nasıl anlaşıldığı bilinebilir. İşte günlük hayattaki bu uygulamanın bilgisayar dünyasındaki karşılığı otomatik metin sınıflandırma işlemidir [2], [21].

Metin arıtma, dokümanların sisteme girmesiyle birlikte denetlenmesi ve daha sonra kullanıcı sorgusuna uygun olanların seçilmesi işlemidir. Metin arıtma, uygun ve uygun olmayan biçimlerde karar verirken aslında dokümanları belirli sınıflara ayırt eder. Bu sebeple Metin arıtma, bir sınıflandırma işlemi olarak da ifade edilebilir [9], [15].

Birçok alanda yeni metinlerin sınıflandırılması görevini profesyonel insanlar üstelenir. Metin sınıflandırma, zaman ve para bakımından maliyeti yüksek bir işlemdir. Bu sebeple otomatik metin arıtma ve sınıflandırma işlemlerinde, teknolojiler ve uygulamaların gelişimi için oldukça büyük bir ilgi vardır. Bağlanım modelleri, en yakın komşu sınıflandırıcıları, karar ağaçları, Bayesian sınıflandırıcıları, kural öğrenme algoritmaları ve yapay sinir ağları gibi pek çok istatistiksel, matematiksel ve otomatik öğrenme teknikleri bu ilgiden dolayı metin sınıflandırma için oluşturulmuştur.

1.6.2.2. Metin Madenciliğinin Ön Aşamaları ve Sınıflama

Metin madenciliği ve metin erişimiyle ilgili olan bütün tekniklerin, kullandıkları ortak yöntemler vardır. Bu bölümde bahsi geçen yöntemler ele alınacaktır.

1.6.2.2.1. Ayrıştırma

Metin veri madenciliğinde yapılan ilk işlem, karakter dizileri olan metinlerin öğrenme algoritmaları ve sınıflandırma işlemleri için uygun duruma getirilmesidir. Bu sebeple bir metin belgesi üzerinde işlem yapılmadan önce bir temizleme ve ayrıştırma işlemi uygulanır. Bu işlem Web sayfaları üzerinde ise, yapılması gereken ilk işlem XML (Extensible Markup Language) ve HTML (Hyper Text Markup Language) gibi her türlü etiketlerin metinden çıkarılmasıdır. Ardından, tüm harfler küçük harfe çevrilir ve belgede yer alan noktalama işaretleri çıkarılır. Daha sonra harf olmayan karakterlerin yerine boşluk karakteri yerleştirilir ve tek harfli sözcükler silinir. Gerekli temizlemeler yapıldıktan sonra son işlem olarak da, belge boşluklara göre kelimelere ayrılır [15].

1.6.2.2.2. Durdurma Kelimelerinin Çıkarılması

Her dilde çok sık kullanılan fakat kendi başlarına bir anlamları olmayan (ve, sonra, ile... gibi) kelimeler veya anlamsal olarak bir ayırıcılığı olmayan kelimeler, durdurma listesi adı verilen bir listede toplanır ki bu listede yer alan kelimeler sınıflandırma işleminde herhangi bir etkiye sahip değildir. Bu sebeple, metin içerisindeki her kelimenin bu listede olup olmadığı incelenir ve bunlar da elenir. Bu liste bir adres hesaplama tablosunda (hash table) da tutulabilir.

1.6.2.2.3. Gövdeleme

Durdurma kelimelerinin çıkarılmasının ardından, her kelimenin eklerinin çıkarılmasıyla kelime kökleri bulunur. Kelime köklerinin bulunması, kelimelerin biçimsel benzerlerinin tespit edilmesi anlamına gelir. Böylece, koşucular, koşucu, koşmak, koş, koşuyorum gibi aynı kök grubundaki kelimeler bir araya getirilmiş olur. Fakat, kök bulmada karşılaşılabilecek iki sorun vardır [11], [19]. Birincisi, bu işlemde çok ileri giderek birbirinden anlamca çok farklı kelimelerin aynı anlam ve kök grubuna bağlanmasıdır. Bu durumda sistem, konuya uygun olmayan dokümanları da konuyla ilgiliymiş gibi yorumlayabilir. Diğer bir sorun da, kelime köklerine ulaşılmaya çalışırken çok az ekin çıkarılmasıdır. Bu durumda da sistem konuya uygun dokümanları, "uygun olmayan" dokümanlar olarak algılayabilir.

Gövdelemeye yarayan birçok farklı algoritma vardır. Bu yöntemlerden biri de tüm dizin sözcüklerinin ve köklerinin Tablo 1.' deki gibi bir tabloda tutulmasıdır.

Tablo 1. Bir kelimenin farklı kullanımları ve kökleri

Yapar	Yapmak
Yapılır	Yapmak
Yapabilir	Yapmak
Yapılırsa	Yapmak
Yaparsa	Yapmak
Yapabilmek	Yapmak

Bu yöntemin dezavantajı, çok fazla saklama alanına gereksinim duyması ve böyle bir tablonun oluşturulmasının zorluğudur. Diğer bir yöntem de, eldeki dokümanlardan oluşturulan bir sözlüğün içindeki her kelimenin, her harfinin tek tek ele alınarak ardıl farklılıklarının incelenmesiyle yapılır. Kökü bulunacak kelimenin sözlük içinde farklı bir kelime olarak bulunabilen ilk n harfi, kelimenin kökü olarak alınır. Mesela sözlüğün içerisinde koş ve koşucu kelimeleri olduğunda Koşucu kelimesinin kökünü bulmak için, k, ko, koş kelimelerine bakılır. Koş sözcüğünün, sözlükte bir kelime olarak görülmesiyle kelimenin kökü bulunmuş olur.

Yukarıdaki yöntemler her dil için geçerli olan yöntemlerdir. Veri kümesi İngilizce metinlerden oluşan çalışmalarda, Porter Stemmer algoritması, daha basit, hızlı olmasına ve performans bakımından diğerleriyle farkı olmaması nedeniyle, bu konu için en sık başvurulan algoritmadır.

1.6.2.2.4. Metin Gösterimi

Metinler sayısal ortamlarda depolanır ancak metin halinde depolanan dokümanlar üzerinde sayısal hesaplamaya dayalı işlemlerin yapılması imkansız olduğu için, dokümanlar farklı gösterim biçimlerine çevrilir. Aşağıda bu gösterim şekillerinden biri olan vektör uzayı modeli anlatılmıştır.

1.6.2.2.5. Vektör Uzayı Modeli

Bu konudaki en çok kullanılan model vektör uzayı modelidir. Bu modele sahip bir dokümanlar kümesindeki her doküman M*N kelime vektörleriyle temsil edilir. M tüm dokümanlardaki her bir farklı kelime sayısını temsil ederken N de elde bulunan tüm dokümanların sayısını temsil eder. Ayrıca, bu vektördeki her bir nitellik, herhangi bir kelimenin o dokümandaki kullanılma sıklığını belirtir. Bir örnek olarak $A= (a_{ij})$, ifadesinde A bir doküman matrisi, a_{ij} ise dokümanlar topluluğundaki her kelimenin içinde bulunduğu bir sözlükte, i numaralı sırada bulunan kelimenin, j numaralı dokümandaki ağırlığını gösterir. Bu yöntem, modern bilgi erişiminin babası olarak kabul edilen Gerard Salton tarafından bilim dünyasına tanıtılmıştır [22].

Metin sınıflama sistemleri bir dokümana ait kelimelerin frekanslarını kullanır. Eğitim kümesi elemanları içerisinden çeşitli ağırlıkları bulur ve bu ağırlıkları sisteme yeni giren dokümanların kategorilerini bulmak için kullanır. Vektör uzayı modelinde, yazılışları aynı fakat farklı anlamlara gelen kelimelerin sorun oluşturabileceği düşünülebilir. Mesela "yüz" kelimesi, "yüzme", "100" veya "insan yüzü" anlamlarına gelecek şekilde kullanılabilir. Ancak bu mesele metin sınıflama tekniklerinde yoktur. Bunun sebebi; sistem, ağırlıkları belirlerken, gerekirse "yüz" kelimesinin ağırlığını düşünür ve diğer kelimelerin ağırlıklarını artırılır.

1.6.2.2.6. Boyut Küçültme

Her kelime, her dokümanda geçerli olmadığı için, yukarıda A ile gösterilen matris genellikle seyrek matristir. Matristeki satır sayısı M, sözlükteki kelime sayısına eşit olduğu için M çok büyük sayıya denk olabilir. Bu durum da matrisin büyümesine ve işlemler sırasında gereksiz zaman ve iş kaybına neden olur. Bu problemi aşmak için farklı sınıflandırma algoritmaları uygulanabilir [1], [21].

1.6.2.2.7. Özellik Seçimi

Bütün boyut küçültme algoritmalarında, tüm dokümanlardaki kelimeler bir sözlük içinde toplanır. Daha sonra küçültme algoritmalarından çıkan sonuçlara göre bu sözlükteki bazı kelimeler çıkartılır. Son aşamada da eldeki dokümanlar tekrar gözden geçirilir ve sadece sözlükte bulunan kelimeler kullanılır.

1.6.2.2.8. Doküman Frekans Eşikleme

Bir kelimenin doküman frekansı, o kelimenin geçtiği doküman sayısına eşittir. Doküman frekans eşitlemeyle her kelimenin doküman frekansı bulunur ve belirli bir sayının altında doküman frekansına sahip olan kelimeler sözlükten çıkarılır. Bu yöntem,

dokümanda belirli bir sayının altında bulunan kelimelerin kümede etkin bir role sahip olmadığı ve kategori belirlemede yetersiz olduğu fikrinden kaynaklanır. Bu yöntem, belirli bir sayı altında dokümanda geçen kelimelerin kümede belirleyici bir röle sahip olmadığı ve kategori belirlemede yetersiz olduğu fikrine dayanır [1], [30].

1.6.2.2.9. Bilgi Kazanımı Yöntemi

Bu yöntem, her bir kelimenin, varlığının veya yokluğunun, kategori seçimi üzerinde etkili olduğu düşüncesinden hareketle ortaya çıkmıştır. $\{C_1 ... C_j\}$ dokümanın ait olabileceği muhtemel kategoriler olabileceği düşünüldüğünde kelime w'nin bilgi kazanım değeri IG(w), aşağıdaki formül ile hesaplanır.

$$IG(w) = -\sum_{j=1}^{k} p(c_j)\log P(c_j) + p(w)\sum_{j=1}^{k} P(c_j|w)\log P(c_j|w) + P(\overline{w})\sum_{j=1}^{k} P(c_j|\overline{w})\, log P(c_j|\overline{w}) \quad (15)$$

Bu ifadeye göre $P(c_j)$ değeri bir dokümanın tüm kategoriler içinde c_j sınıfına ait olma ihtimali ve P(w) topluluktaki tüm dokümanlar içinde bir dokümanın içerdiği w kelimesinin bulunma olasılığıdır. Burada $P(c_j|w)$ ve $P(c_j|\overline{w})$ değerleri sırasıyla, c_j sınıfındaki dokümanlardan en az bir kere w kelimesinin görünme ihtimali ve c_j sınıfındaki dokümanlardan birinde hiç w kelimesinin görünmemesi ihtimali olarak tanımlanmaktadır.

Bilgi kazanımı değeri, koleksiyondaki her eğitim doküman kelimesi için hesaplanarak belli bir değerin altında olan kelimeler koleksiyondan çıkarılır.

1.6.2.2.10. Ağırlıklandırma

Yukarıda belirtilen A matrisinin taşıdığı ağırlık değerlerinin tespitinde birçok yöntem uygulanabilir. Fakat bu yöntemlerin hemen hemen hepsi iki önemli hususa dayanır;

Bir sözcük, bir dokümanın içinde ne kadar sık tekrarlanırsa, o dokümanın bir kategoriye gönderilmesinde o kadar etkili olur.

Bir sözcük ne kadar çok farklı dokümanda yer alıyorsa, o sözcüğün ayırt edici özelliği o kadar azalır. Aşağıda kısaca açıklanacak olan bu yöntemde kullanılan temel değişkenler f_{ij} , i indeksli kelimenin j metni içerisindeki kullanım sayısını, N toplam doküman sayısını, M toplam sözcük sayısını, n_i bu sözcüğe sahip olan dokuman sayısını ifade eder.

1.6.2.2.10.1. Boole Değerler ile Ayırma

Boole ayırma yönteminde en sade ve basit olan yaklaşım, eğer kelime dokümanda mevcutsa ağırlık değerini 1'e, eğer yoksa 0'a eşitler.

$$a_{ij} = \begin{cases} 1 & eğer\ f_{ij} > 0 \\ 0 & diğer\ durumlar \end{cases} \quad (16)$$

1.6.2.2.10.2. Kelime Frekans Ağırlıklandırma

Bu yöntem kullanılan basit yöntemlerden biridir. Bu yöntemde, diğeri olan kelime frekans ağırlığında, ağırlık kelimenin doküman içerisindeki ham frekansına eşitlenir.

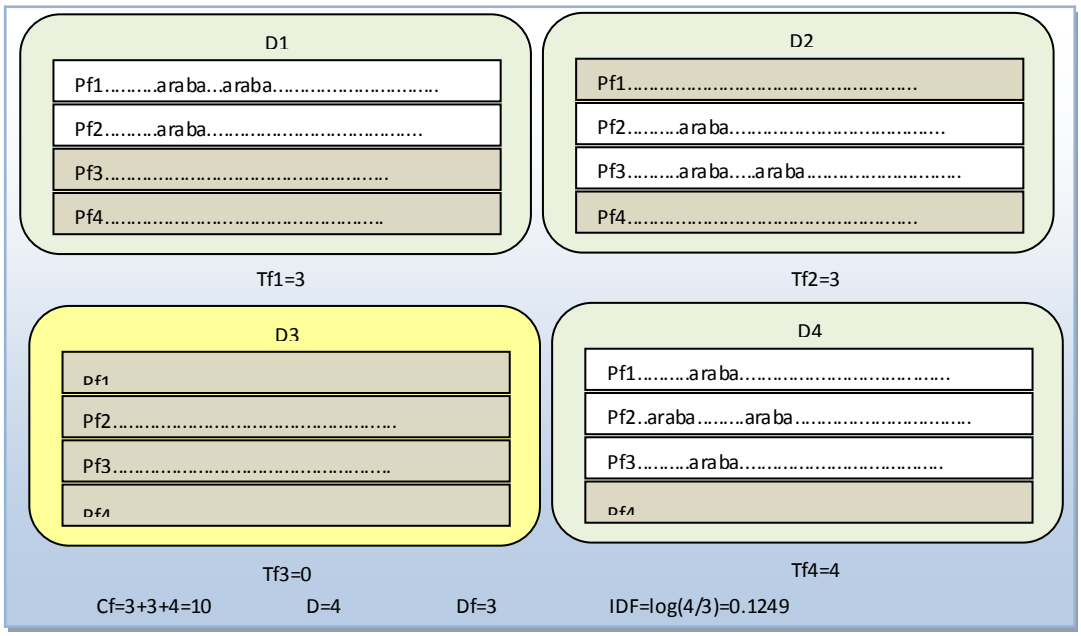
$$a_{ij} = f_{ij} \quad (17)$$

1.6.2.2.10.3. tf * idf Ağırlıklandırma

Yukarıdaki iki ağırlıklandırma yönteminde de sözcüğün tüm dokümanlar içerisindeki etkileri dikkate alınmadan ağırlık değerini tespit ediliyordu. Tf * idf (Term Frequency * Inverse Document Frequency) yönteminde ise bu iki yöntemin aksine tüm metinler göz önünde bulundurularak ağırlandırma yapılır. Bu yöntem, eğer bir kelime az sayıda dokümanda yer alıyorsa, kelimenin o dokümanın kategorisinin belirlenmesinde önemli olduğu, eğer bir kelime çok sayıda dokümanda kullanılmamışsa, kelimenin ayırt edici gücünün az olduğu düşüncesine sahiptir. Bu yöntem, eğer bir kelime az sayıda dokumanda geçiyorsa, kelimenin o dokumanın kategorisinin belirlenmesinde önemli olduğu, eğer bir kelime çok sayıda dokumanda kullanılıyorsa, kelimenin ayırt edici gücünün az olduğu fikriyle açıklanabilir. Yeni ağırlık değeri, aşağıdaki formülle hesaplanır.

$$a_{ik} = f_{ik} * \log\left(N/n_i\right) \tag{18}$$

Şekil 6'de tf *idf ağırlıklandırma yönteminde hesaplanma nasıl yapılabilir göstermektedir;



Şekil 6. Küme üzerinde "araba" kelimesinin Tf*İdf ağırlıklandırma yöntemine göre ağırlandırlması [21].

Şekil 6 de beyaz, yeşil, gri, sarı ve mavi yuvarlaklar sırasıyla, araba kelimesini içeren terim, araba kelimesini içeren doküman, araba kelimesini içermeyen terim, araba kelimesini içermeyen doküman ve dokümanların kümesi olarak belirlenmektedir.

1.6.2.2.10.4. tfc-Ağırlıklandırma

Uzun dokümanlar daha fazla sözcük içerdiklerinden, bu dokümanlarda çok sayıda farklı sözcüğün kullanılması ve bu sözcüklerin frekanslarının da kısa dokümanlara göre daha fazla olasılığı yüksektir. İşte tfc (Term Frequency Component) ağırlıklandırma tf * idf nin bu olasılığını dikkate alınarak yapılan bir düzgeleme işlemidir [1].

tf * idf formülünde görülen, kelimenin tüm dokümanlarda tekrar etme sayısı sözlük tablosundan ve her bir dokümandaki kelimenin kullanılma sayısı ise doküman vektöründen alınmıştır.

$$a_{ik} = \frac{f_{ik} * log\left(\frac{N}{n_i}\right)}{\sqrt{\sum_{j=1}^{M}\left[f_{jk} * log\left(\frac{N}{n_j}\right)\right]^2}} \tag{19}$$

1.6.2.2.10.5. ltc Ağırlıklandırma

1tc (Logarithmic Term Component) ağırlıklandırma yöntemi, tcf'nin biraz daha değiştirilmiş hali olup ham frekanslar yerine logaritma kullanarak, frekanslardaki büyük çaplı değişiklilerin etkisini azaltır.

$$a_{ik} = \frac{\log(f_{ik}+1) * log\left(\frac{N}{n_i}\right)}{\sqrt{\sum_{j=1}^{M}\left[\log\left(f_{jk}+1\right) * log\left(\frac{N}{n_j}\right)\right]^2}} \quad (20)$$

1.7. Model Performansını Değerlendirme

Model performansını değerlendirmek için kullanılan temel kavramlar kesinlik, duyarlılık, hata oranı ve F-ölçütüdür. Modelin performansı, doğru sınıfa atanan test sayısı ve yanlış sınıfa atılan test sayısı miktarlarıyla ilgilidir. Değerlendirme ölçümlerini hesaplamak için aşağıdaki 4 kavramın tanımlanması gerekir;

a: TP (True Pozitif)
b: FN (False Negatif)
c: FP (False Pozitif)
d: TN (True Negatif)

TP ve TN değerleri doğru sınıflandırılmış örnek sayısıdır. False Pozitif (FP), aslında 0 (negatif) sınıfındayken 1 (pozitif) olarak tahmin edilnmiş örneklerin sayısıdır. False Negative (FN) ise 1 (pozitif) sınıfındayken 0 (negatif) olarak tahmin edilmiş örneklerin sayısını ifade eder [24].

Tablo 2. İki sınıflı bir veri kümesinde oluşturulmuş modelin karışıklık matrisi [24].

		Öngörülen Sınıf	
Doğru Sınıf		Sınıf=1	Sınıf=0
	Sınıf=1	a	b
	Sınıf=0	c	d

1.7.1. Doğruluk – Hata Oranı

Modele ait doğruluk oranıdır ve en popüler, basit ve belirleyici ölçüttür ve bundan dolayı yaygın bir şekilde kullanılmaktadır. Doğru sınıflandırılmış test sayısının (TP +TN), toplam test sayısına (TP+TN+FP+FN) oranıdır [7].

$$doğruluk = \frac{TP + TN}{TP + FP + FN + TN} \tag{21}$$

Hata oranı, yanlış sınıflandırılmış test sayısının (FP+FN), toplam test sayısına (TP+TN+FP+FN) oranıdır.

$$hata\ oranı = \frac{TP + TN}{TP + FP + FN + TN} \tag{22}$$

1.7.2. Kesinlik

Kesinlik, sınıfı 1 olarak tahmin edilmiş Pozitif test sayısının, sınıfı 1 olarak tahmin edilmiş tüm test sayısına oranıdır [7], [22]:

$$kesinlik = \frac{TP}{TP + FP} \tag{23}$$

1.7.3. Duyarlılık

Doğru sınıflandırılmış pozitif test sayısının toplam pozitif test sayısına oranıdır [7], [20].

$$duyarlılık = \frac{TP}{TP + FN} \tag{24}$$

1.7.4. F-Ölçütü

F-ölçütü, kesinlik ve duyarlılıktan elde edilen orandır [7], [20].

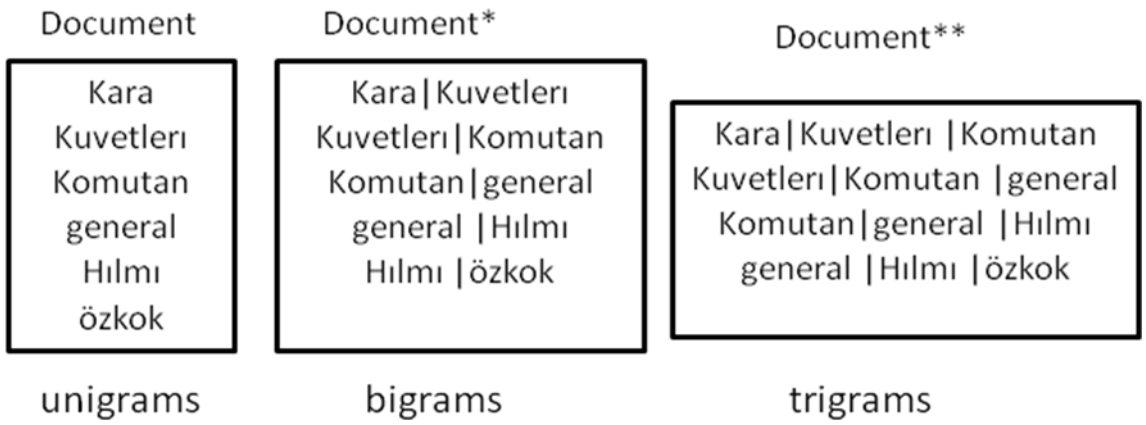
$$F1 = \frac{2.kesinlik.duyarlılık}{kesinlik + duyarlılık} \tag{25}$$

2. YAPILAN ÇALIŞMALAR

2.1. Ön İşlem Aşamaları

2.1.1. Metinlerin Çözümlenmesi

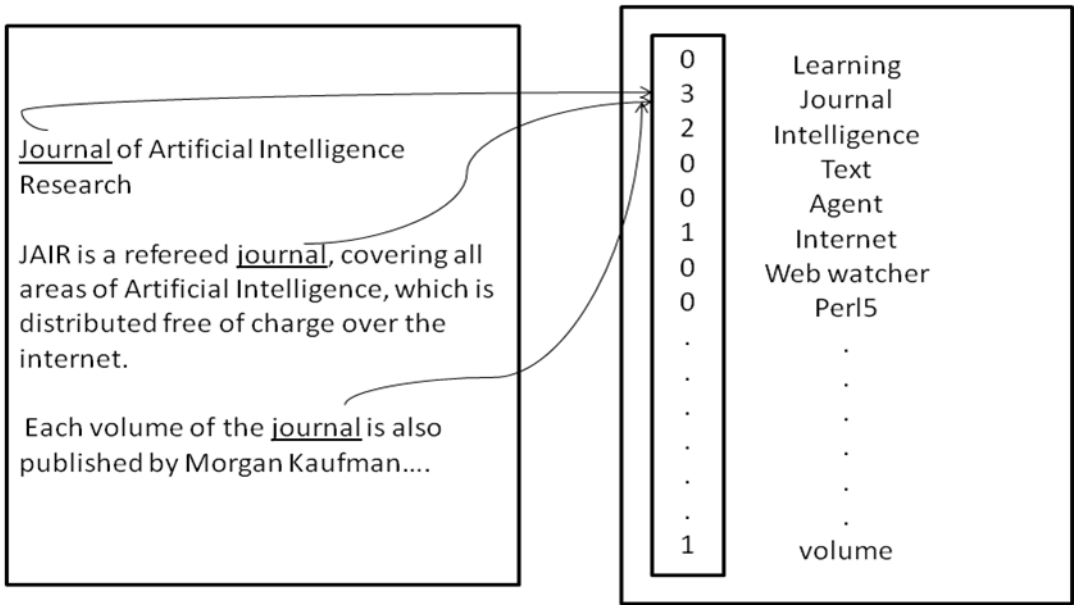
İnternetten elde edilen metinler, çok farklı şekillerde olabilir. Bu metinlerin sınıflandırma için uygun bir temsile (gösterimini) veya biçime dönüştürülmesi gerekir. Bunun için öncelikle HTML veya diğer etiketleri kaldırılarak ve sade metin elde edilir ve daha sonra sayısal ve tüm noktalama ve diğer tüm işaretleri kaldırma, tokenaize etmek veya simgeleştirme, standart formlara çevirmek ve stopwords çıkarmak gibi işlemler uygulanarak elde edilen metin, metin madenciliği için hazır hale gelir. Ayrıca, metinleri yaratan kelimeler farklı ekler alarak, değişik anlamlarda kullanılmaktadır. Morfolojik analiz veya stemming ile doküman içindeki her bir kelimenin tüm morfolojik halleri bulunur ve kelimelerin kökleri çıkartılır. Ancak, bu çalışmada iki farklı dili ele alındığı için eğitim kümesi ve test kümesi küçük boyutlarda seçilmiştir. Bu nedenle dolayı morfolojik analiz bu çalışmada kullanılmamış ve N-gram tokenize kullanılmıştır. N-Gram tokenin boyutu N in boyutu ile sınırlıdır. Aşağıda şekil 7'da N-Gram tokenın boyutlarından 1, 2 ve 3 boyut olan N-Gramlar gösterilmiştir [9], [10].



Şekil **7.** N-Gram boyutlarından 1,2 ve 3 boyut olan N-Gramlar

Bu çalışmada N bir seçilerek unigram olarak kullanılmıştır.

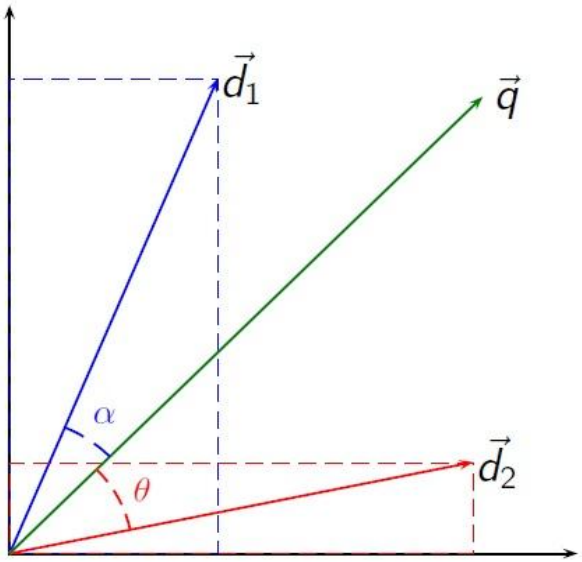
Tokenize yöntemi, belgeleri kelime sayılarına dönüştürmekte ve bu kelimelerin her hangi bir anlama sahip olup olmamasından bağımsızdır. Örnek olarak, "Mary is smarter than John" ve "John is smarter than Mary" bag of word tablosında eşittir. Metinlerin içerdiği kelimelere tokenize olduktan sonra kelimeler standart formlara çevrilir. Yani belgeler içinde tüm büyük harfleri küçük harflere dönüştürülür ve daha sonra sık sık görülen kelimeler kaldırılır. Örneğin İngilizce de (a, the, and…) gibi kelimeler ve Türkçe de (o, bu, ve,…) gibi kelimeler çıkarılır. Daha sonra çok nadir kullanılan kelimeler de kaldırılır. Ardından, metin noktalama işaretlerinden arındırılır. Metin hazırlandıktan sonra belgeyi temsil etmek için iki temel yaklaşım vardır. Bunlar bag of words tablosunu üretmek ve ondan da vektör uzayını üretmektir. Şekil 8' bu yaklaşımın, daha önceden belirlenmiş anahtar kelimelerin, metin içinde aranmasıyla elde edilen değerlerin frekansları ile oluşturulmasını göstermektedir [21], [26].



Şekil 8. Bag of words ile vektor uzayı üretmek

Terimler, Şekil 8'ye göre "*i*" bir dokümanda geçen bir temim geçme sıklığına, "*j*" veri tabanındaki doküman sayısı frekansları ile ağırlıklandırılmıştır. Vektör uzay modeli aşağıdaki örnekle gösterilmiştir. Bu modelde dokümanlar, kelimelerden ulaşan vektörleri ile ifade edilmiştirler. Şekil 9' de w'ler kelimelerin frekanslarını temsil etmektedirler ve D'ler eğitim dokümanlarının vektörünü, Q ise sınıflandırılacak dokümanın vektörünü temsil etmektedir.

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

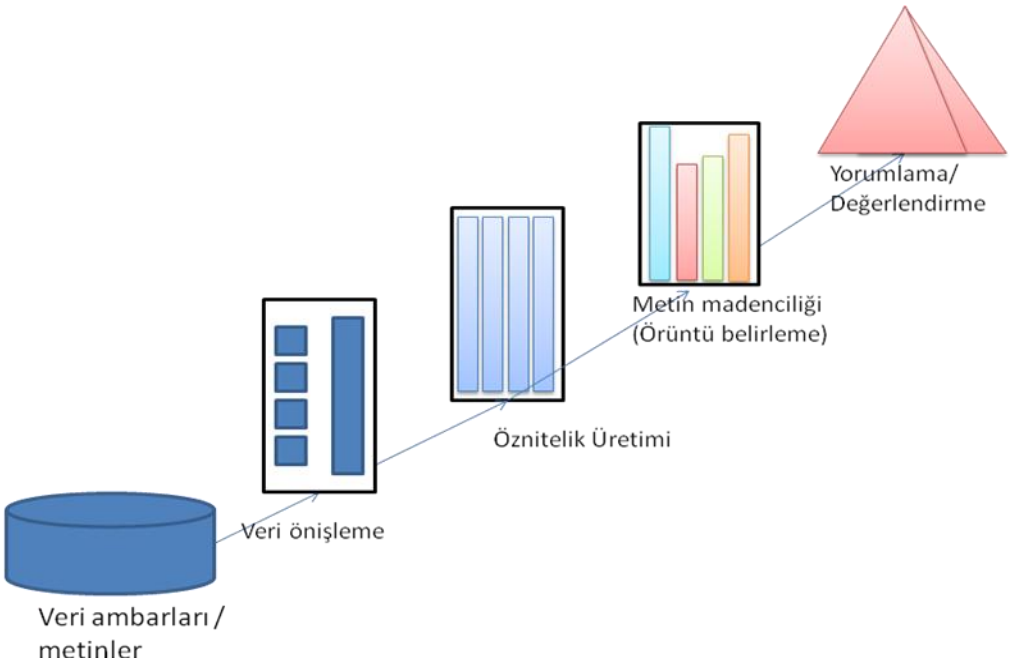


Şekil 9. Vektör uzay modeli

Q vektörü, sınıflandırılacak doküman, *D1 ve D2* vektörleri ile kıyaslanır ve vektörler arasındaki açının kosinüsü hesaplanarak kıyaslama yapılır. 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 vektori d1 le aynı kategoridedir [22].

2.1.2. Geliştirilen Sistemin Açıklanması

Bu çalışmada gerçekleştirilen uygulamanın aşamalar aşağıdaki Şekil'10 de gösterildiği gibi beş adımdan oluşmuştur.



Şekil 10. Geliştirilen sistemin oluşan aşamaları

Yapılan bu çalışmada çevrimiçi (online) beş farklı konuda seçilen Türkçe ve İngilizce metinlerin orijinal dilde analiz edip hangi kategoriye ait olduğunu tespit ettikten sonra aynı metinler bilgisayarlı çevirici (Google translator) kullanılarak diğer dile çevirtilerek tekrar sınıflandırmaktadır. Burada amaç, çevrilmiş metinler üzerinde sınıflandırma yönteminde nasıl performans gösterildiğini ve bilgisayarlı çeviricilerin veri madenciliğinde etkilerini incelemektir. Bu esasa dayanarak Web veya www kapsamında metin sınıflandırma, metin analizi ve çözümlemesi için veri madenciliği için yazılımı geliştirilmiş ve uygulanmıştır.

Şekil 11. Çevirici ve sınıflandırıcı uygulaması

Geliştirilen program eğitim kümesi olarak istenilen istenen sayıda kategori ve dokümana sahip olabilir. Test setindeki dokümanlar, orijinal dilde veya çevrildiği dilde geliştirilen programla sınıflandırılabilir. Geliştirilen uygulama genel olarak aşağıdaki bileşenlerden oluşmaktadır.

1. Web Tarayıcı Modülü
2. Çevrici Araçlar ve Metin Çözümleme Araçları
3. Metin Sınıflandırma Araçları

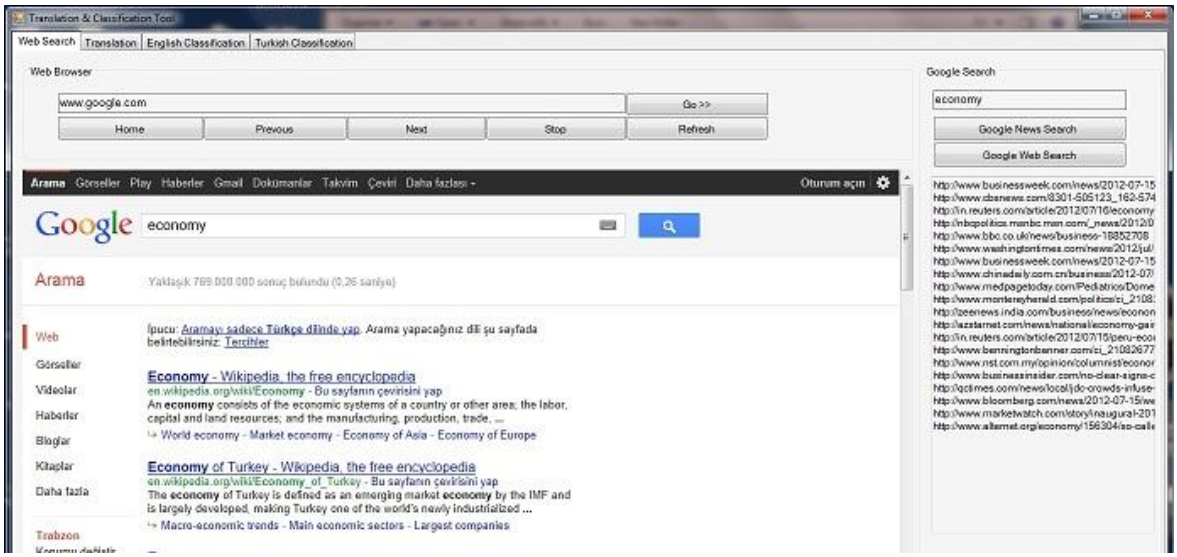
2.1.2.1. Web Tarayıcı Modülü

Bu çalışmada, Web tarayıcısı olarak Google tarama motoru kullanarak haber ve siteler üzerinde tarama ve veri maddeciliği algoritmalarının uygulamaları gerçekleştirilmiştir. Ayrıca, her hangi bir Web adresine ulaşabilmek için tarayıcıya özel tarayıcı uygulaması da geliştirilmiştir. Şekil 11' da gösterildiği gibi tarama dört farklı biçimde yapılabilmektedir. Geliştirilen Web tarayıcı ile istenen Web sayfası indirilebilir. Örneğin, şekil 12' deki gibi Web tarayıcısında "www.google.com" adresi yazılırsa, Google

tarayıcı aracılığıyla istenen her hangi bir Web sayfasına ulaşılabilinir ve elde edilen sayfa geliştirilen yazılımla sınıflandırılabilir.

Şekil 12. Geliştirilen Web Tarayıcı

Öte yandan Google tarayıcı da kullanılarak, yazılan kategorinin ismiyle (Örneğin haber gibi) Google tarayıcısı tarafından veya Web sayfaları tıklandığında elde edilen Web sayfaları sınıflandırma ve ilgili sonuçlar Şekil 13'deki gibi aynı ekranda gösterilebilir.

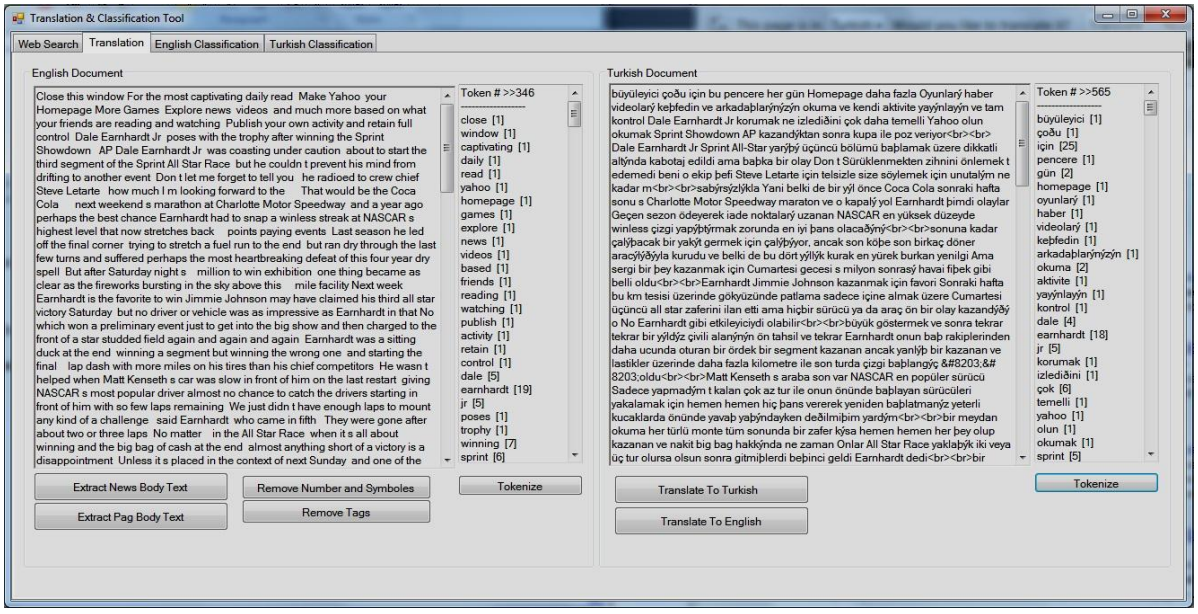


Şekil 13. Tarama ve sonuçları

Bulunan sonuçlardan herhangi birisi üzerine tıklayarak sınıflandırma aracına devredilebilir.

2.1.2.2. Çeviri ve Metin Çözümleme Araçları

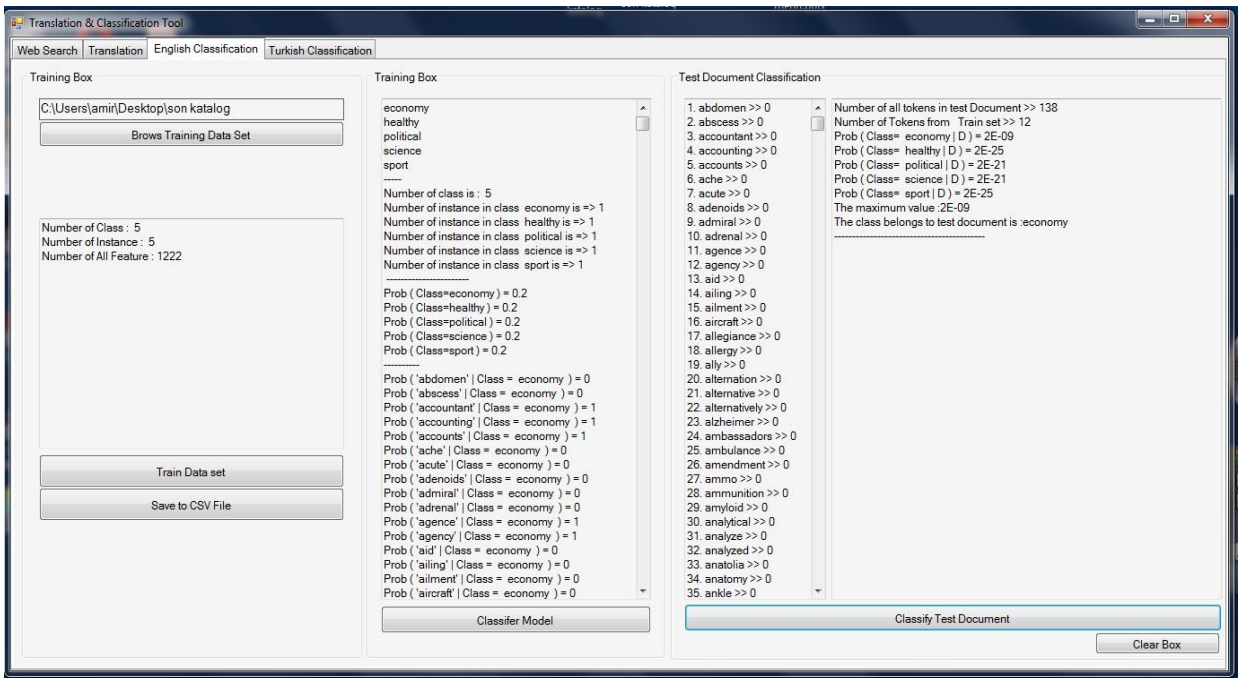
Geliştirilen sistem ikinci olarak Web sayfasının HTML hali çözümler ve etiketleri kaldırır. Böylece, saf bir metin elde edilmiş olur. "Extract news body text" veya "Extract tag body text" butonlar ile etiketler ayıklanmış metinden otomatik olarak sayılar, her türlü noktalama işaretleri ve stopword'ları çıkartılır. Daha sonra Tokenize butonuna tıklanarak metinindeki kelimelere tokenize edilir. Eğer metin İngilizce ise İngilizce ve eğer Türkçe ise Türkçe sonuç üretilir. Metinin tercüme etmek için Şekil 14' de verilen "Translation To Turkish" ve "Translation To English" butonları üzerine tıklanılır. Çeviri, Google translator kullanılarak yapılır. Çevrilen metin aynen orijinal metin gibi çözülür ve tokenize edilir.



Şekil 14. Metin çıktısı ve Çevirisi

2.1.2.3. Metin Sınıflandırma Araçları

Önceki aşamalarda hazırlanan metin, metin madenciliği yöntemleri için uygun hale gelmiştir. Böylece, metin sınıflandırma modülü üzerinde işlem yapılabilir hale dönüştürülmüştür. Şekil 15' de görüldüğü gibi sınıflandırma modülü üç kısımdan oluşur. Bu öncelikle eğitim kümesini (Training Data set) analiz edilir. Eğitim kümesi dinamik olarak belirlemektedir. Çalışmada 6 kategori Ekonomi (Economy), Politika (Policy), Tıp (Medicine), Din (Religion), Sopr (Sport) ve Bilim (Science) kategorileri seçilmiştir. Seçilen her kategori 50 dokümana (Instance) sahiptir. Eğitim kümesi CSV (Comma-Separated Values) dosyası olarak kayıt edilebilir. Böylece, ilgili dosya Weka uygulamalarında kullanılabilir.



Şekil 15. Sınıflandırma modülü

Geliştirilen sitemde sınıflandırma yöntemi olarak Naive Bayes metodu kullanılmıştır. Eğitim kümesi ile eğitim yapıldıktan sonra sınıflandırma modeli bu eğitim kümesine göre hesaplanır ve her kelimenin ihtimali, mevcut olan altı kategori için elde edilir.

"Classify Test Document" butonuna tıklandığında sınıflanırıma için hazırlanan metin dokümanın sınıflandırma modeli ile ait olduğu sınıf belirlenir.

2.1.2.3. Geliştirilen sistemin sonuçlarına göre yapılan değerlendirmeler

Tablo 3. deki değerler 20 test üzerinden elde edilmiştir.

Tablo 3. Geliştirilen Sistemin Sonuçlar

Kategori	Değerlendirmeler
Türkçe	% 80
İngilizce	% 80
Türkçe - İngilizce Çeviri	% 75
İngilizce – Türkçe Çeviri	% 75

2.2. Weka Araçı

2.2.1. Weka Sistemi ve Elde Edilen Sonuçlar

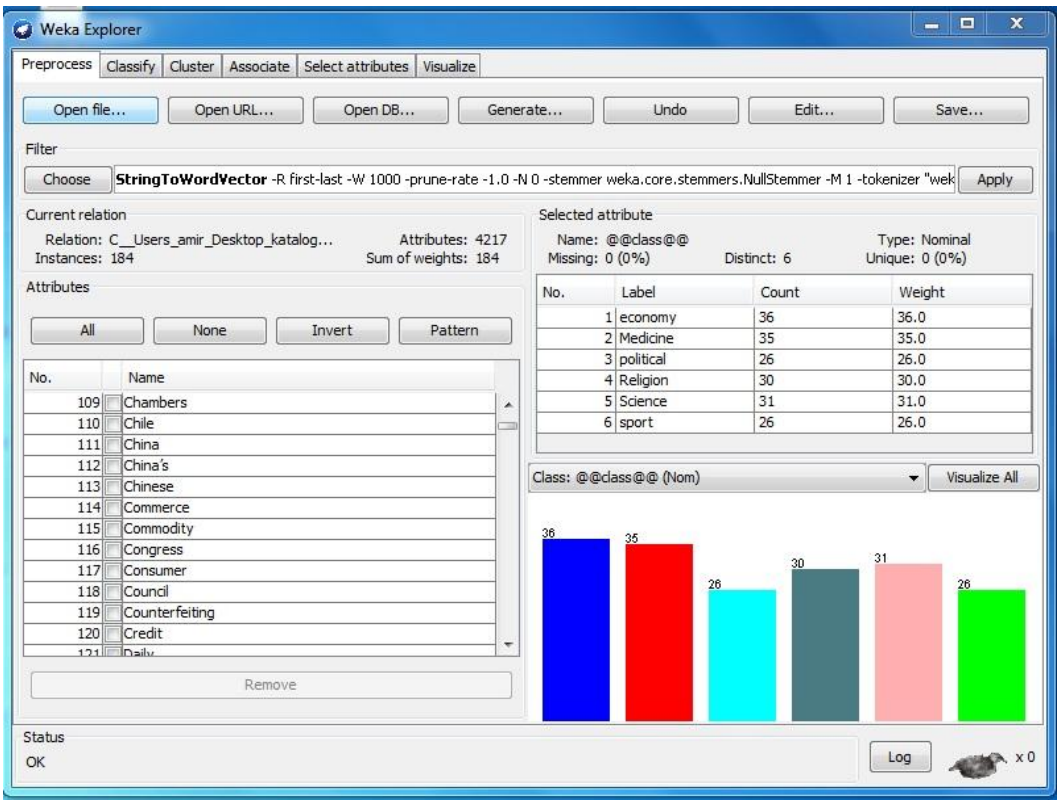
Weka programı Waikato Üniversitesinde Java programlama dilinde geliştirilmiştir. Ücretsiz ve açık kaynak kodlu olan programda pek çok sınıflandırma, regresyon, demetleme, bağıntı kuralları, yapay sinir ağları algoritmaları ve önişleme araçları mevcuttur. Aynı zamanda Weka oldukça yaygın kullanılan bir referans veri madenciliği aracıdır [6]. WEKA, metin tabanlı arff, csv, c45, libsvm, svmlight, Xarff formatlarda olan temel veri kaynaklarını desteklemektedir. Weka, ham verinin işlenmesi, öğrenme metotlarının veri üzerinde istatistiksel olarak değerlendirilmesi, ham verinin ve ham veriden öğrenilerek çıkarılan modelin görsel olarak izlenmesi gibi veri madenciliğinin tüm basamaklarını destekler [31].

Bu çalışmada üretilen dosya formatı CSV olduğu için Weka programını tarafından sınıflandırmalarda işleminde kullanabilir. Yazılımla elde edilen CSV dosyası şekil 16’ de gösteri gibi kelime frekansına bağlı olan bir vektör tablosudur.

e	caesar	cates	catteinate	catteine	cairo	caja	cake	calcium	calendar	calit	california	californiar	caliphate	call	called	callers	calling	calls
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Şekil 16. Kelime frekansına göre üretilmiş vektör tablosu

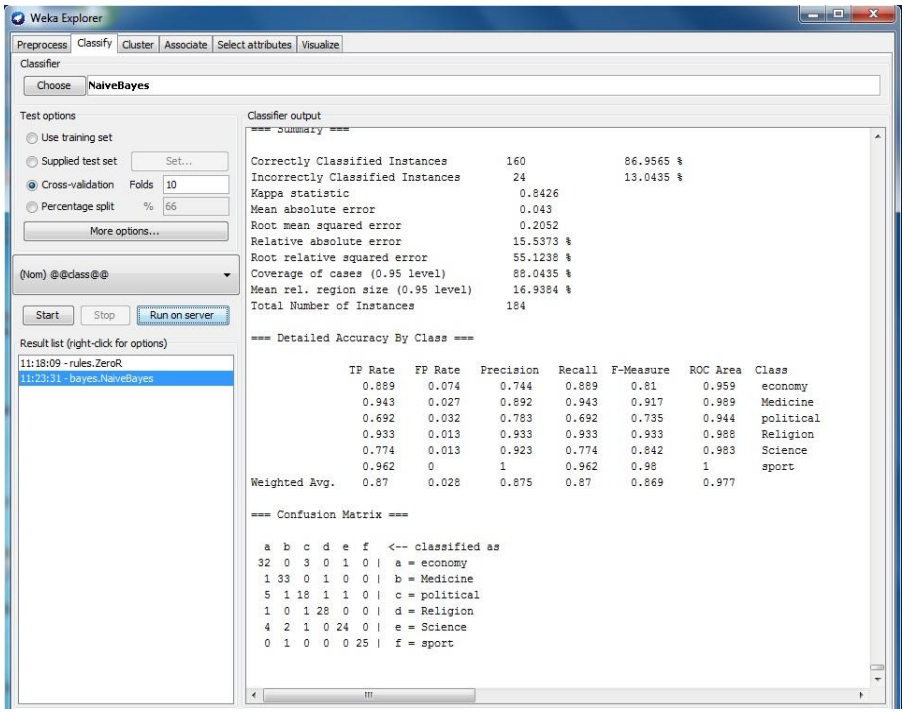
Weka programı oluşturulan vektör tablosunu Weka Explorer arayüzünde olan Preprocess etiketi ve Open file komutlarıyla açıldıktan sonra kullanılacak işlemler için ilgili sekmelere tıklanır.



Şekil 17. Weka programının kullanılanımı ve Explorer arayüzü

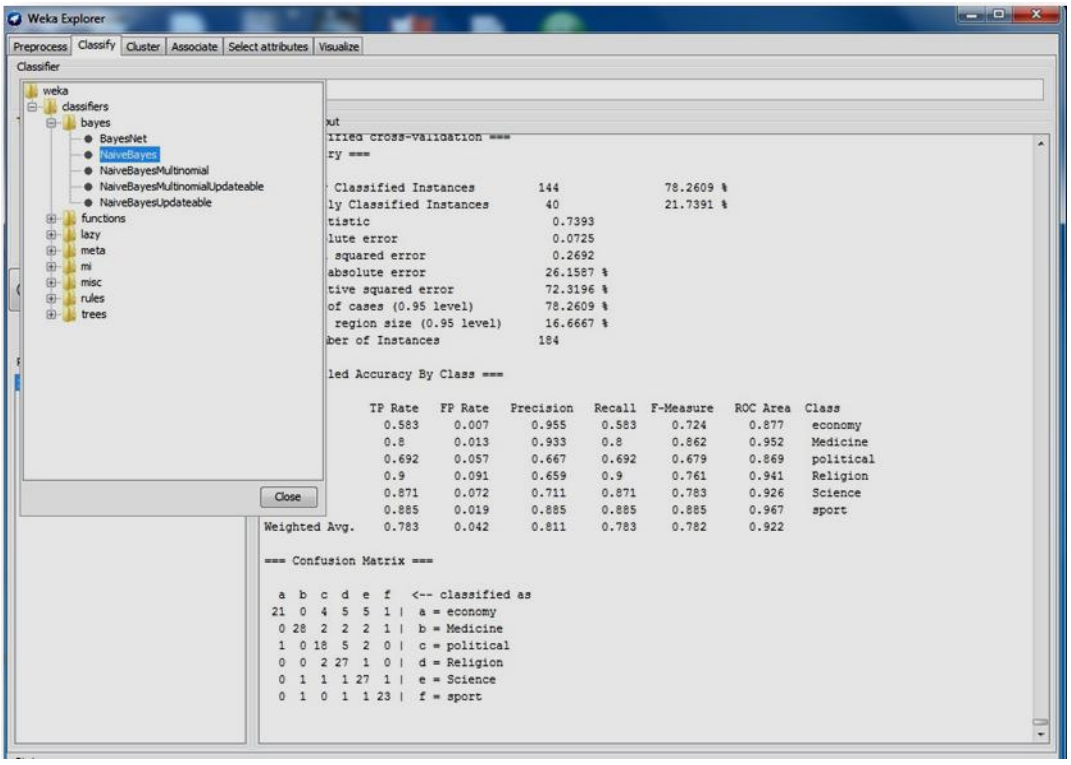
2.2.1.1. Weka Programıyla Doküman Sınıflandırma

Weka programı tarafından kullanılacak hale getirilen metine rahatlıkla sınıflandırma işlemleri uygulanabilir. Sınıflandırılacak dosya yüklendikten sonra Şekil 18 ve 19' de sergilendiği gibi sınıflandırma işlemleri gerçekleştirilir.



Şekil 18. Weka programın sınıflandırma arayüzü

Sınıflandırıcı (Classifier) içinde olan seçme (choose) butonuna tıklanırsa sınıflandırma algoritmalarının listesi görülebilir. Bu listeden istenen sınıflandırma algoritmasını seçerek ilgili dosya sınıflandırılır.



Şekil 19. Weka sınıflandırma listesi

Bu aşamada seçeneğinde (Test options) dört seçenek vardır. Bunlar;

1. Eğitim Seti Kullan (Use training set): eğitim kümesi ve test kümesi aynıdır.
2. Test Setini Kullan (Supplied test set): eğitim seti open file den açılan dosya ve test seti, set düğmesi tıklandıktan sonra gelen pencereden elde edilir.
3. Çapraz Doğrulama (Cross Validation): folds da yazılan adede göre veri seti n gruba ayrılır, bir grup test için ayrılırken geriye kalan n-1 grupla sistem eğitilir.
4. Yüzdelik Dilim (Percentage Split): eğitim seti ve test seti yüzde olarak bölünür ve sınıflandırılır.

3. BULGULAR VE İRDELEME

Bu çalışmada, gerçeklenen yazılımla ve Weka programı ile metinsel madencilik ve metinsel madencilikte bilgisayarlı çeviricilerin kullanımına yönelik sonuçlar elde edilmiştir. Gerçeklenen yazılım ile yapılan uygulamalarda 6 kategoride eğitim kümesi seçilmiş ve her kategori 30 dokümanla eğitim yapılmıştır. Test dokümanı olarak ilgili kategorilerde herhangi bir Web sayfası veya herhangi metin seçilebilir. Geliştirilen yazılımda Naive Bayes sınıflandırma yöntemi kullanılmıştır.

Yapılan uygulamalarda elde edilen sonuçlara göre çevrilen dokümanların için elde edilen sınıflandırma performansının orijinal dokümanların sınıflandırma sonuçlarına göre küçük bir aranda düştüğü görüşmüştür. Bu durumun birinci nedeni olarak, kullanılan bilgisayarlı çeviricin tüm kelimeleri çevirememesinden veya yanlış anlama çevrilmesinden kaynaklandığının öngörmekteyiz.

Tablo 4**.** Kategoriler ve doküman sayıları

Doküman Kategorileri	**Türkçe Doküman Sayıları**	**İngilizce Doküman Sayıları**
Spor	153	153
Ekonomi	153	153
Tıp	153	153
Politika	285	285
Bilim	153	153
Din	100	100
Toplam Doküman Sayısı	997	997

Weka programını kullanarak da benzer sonuçlar elde edilmiştir. Yani çevrilen dokümanların sınıflandırma performanslarının orijinal dokümanların sınıflandırmalarına göre benzer olarak küçük bir yüzde olarak düşük olduğu görülmüştür.

Weka sınıflandırmalarda Naive Bayes (NB), Multinomial Naive Bayes (MNB), Sequential Minimal Optimization (SMO) ve J48 algoritmalar kullanılmıştır. Elde edilen sonuçlara göre Multinomial Naive Bayes en iyi performansa sahip ve J48 en düşük performansea sahip olduğu belirlenmiştir. Elde edilen sonuçlar aşağıdaki Tablolarda göstermektedir.

Weka ile yapınla çalışmalarda da 6 kategoride (Ekonomi, Tıp, Politika, Din, Bilim ve Spor) sınıflandırma yapılmıştır. Politika ve Din kategorileri haricindeki bütün kategorilerde 153 doküman kullanılmıştır. Politikada 284 ve Din kategorisinde 100 doküman bulunmaktadır.

Tablo.4 ‘de kategoriler ve doküman sayıları gösterilmiştir. Dokümanlar ve kategoriler Türkçe ve İngilizce için aynı sayıda seçilmişlerdir.

Çalışmada ve testlerde kullanılan 997 dokümanın %66 eğitim kümesi ve %34 test kümesi olarak kullanılmıştır. Bu durumda İngilizce metinlerin sınıflama performansı Tablo 5’de gösterilmiştir.

Tablo 5. İngilizce Metinlerin Sınıflaması (Percentage split = 66% kullanılarak)

Kategori	SMO	NBM	NB	J48
Spor	0.898	0.937	0.897	0.611
Ekonomi	0.935	0.954	0.947	0.805
Tıp	0.889	0.937	0.897	0.717
Politika	0.932	0.972	0.916	0.694
Bilim	0.855	0.925	0.89	0.722
Din	0.948	0.988	0.962	0.866
Ortalama	0.912	0.944	0.935	0.777

Yine aynı metinler üzerinde Cross- validation Folds = 10'e göre aynen testler yapılmış ve Tablo 6'da verilen sonuçlar elde edilmiştir. Burada dokümanlar 10 kısma bölünüş ve 9 kısmı eğitim kümesi ve 1 kısmı test kümesi olarak kullanılmıştır.

Tablo 6. İngilizce Metinlerin Sınıflaması (Cross- validation Folds = 10 kullanılarak)

Kategori	SMO	NBM	NB	J48
Ekonomi	0.875	0.914	0.9	0.693
Tıp	0.954	0.957	0.947	0.857
Politika	0.914	0.926	0.908	0.755
Din	0.927	0.941	0.885	0.828
Bilim	0.899	0.929	0.92	0.711
Spor	0.961	0.977	0.967	0.873
Ortalama	0.921	0.939	0.921	0.78

Yine aynı dokümanlar bilgisayarlı çevirici (Google Translator) kullanılarak Türkçeye çevrilmiş ve ayni şekilde aynı yöntemleri kullanarak sınıflandırılmıştır. Elde edilen sonuçlar aşağıdaki Tablo 7 ve 8 'de verilmiştir.

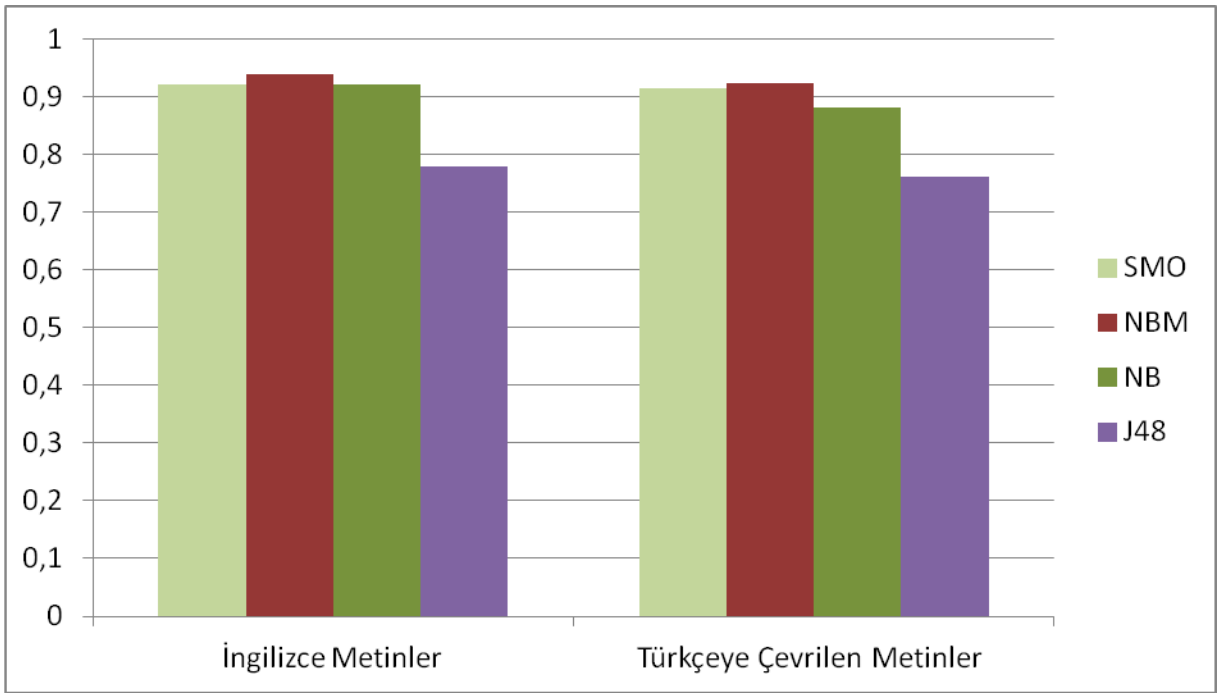
Tablo 7. İngilizce dokümanların Türkçeye çevrildikten sonra sınıflama sonuçları (Percentage split = 66% kullanılarak)

Kategori	SMO	NBM	NB	J48
Ekonomi	0.851	0.867	0.848	0.519
Tıp	0.966	0.957	0.931	0.842
Politika	0.891	0.921	0.878	0.652
Din	0.889	0.854	0.771	0.737
Bilim	0.898	0.876	0.86	0.695
Spor	0.95	0.97	0.94	0.824
Ortalama	0.909	0.912	0.877	0.711

Tablo 8. İngilizce dokümanların Türkçeye çevrildikten sonra sınıflama sonuçları (Cross-validation Folds = 10 kullanılarak)

Kategori	SMO	NBM	NB	J48
Ekonomi	0.867	0.885	0.857	0.688
Tıp	0.947	0.947	0.916	0.817
Politika	0.906	0.926	0.871	0.718
Din	0.918	0.879	0.775	0.821
Bilim	0.892	0.903	0.874	0.75
Spor	0.968	0.98	0.964	0.828
Ortalama	0.915	0.923	0.881	0.761

İngilizce metinler ve çeviriler üzerindeki kıyaslamalar aşağıdaki şekilde göstermektedir.



Şekil 20. İngilizce metinlerin sınıflaması ve Türkçeye çevrilen metinlerin sınıflamalarının kıyaslama

Bu uygulamada Türkçe dokümanların sınıflandırılması ve daha sonra Türkçe den İngilizceye çevrilen metinlerin sınıflandırılması sonuçları Tablo 9, 10 ve Şekil 21'de verilmiştir.

Tablo 9. Türkçe Metinlerin Sınıflaması (Cross- validation Folds = 10 kullanılarak)

Kategori	SMO	NBM	NB	J48
Ekonomi	0.838	0.915	0.875	0.75
Tıp	0.747	0.887	0.835	0.673
Politika	0.808	0.878	0.82	0.698
Din	0.701	0.809	0.716	0.644
Bilim	0.788	0.862	0.782	0.698
Spor	0.922	0.96	0.912	0.776
Ortalama	0.807	0.888	0.83	0.69

Tablo 10. Türkçe Metinlerin Sınıflaması (Percentage split = 66% kullanılarak)

Kategori	SMO	NBM	NB	J48
Ekonomi	0.76	0.866	0.8	0.66
Tıp	0.744	0.875	0.829	0.627
Politika	0.735	0.859	0.793	0.653
Din	0.677	0.738	0.744	0.522
Bilim	0.782	0.836	0.777	0.631
Spor	0.928	0.99	0.904	0.74
Ortalama	0.77	0.865	0.808	0.644

Türkçe metinlerin İngilizceye çevrilmesinden sonra sınıflandırılmaları sonuçları Tablo 11 ve 12'de verilmiştir.

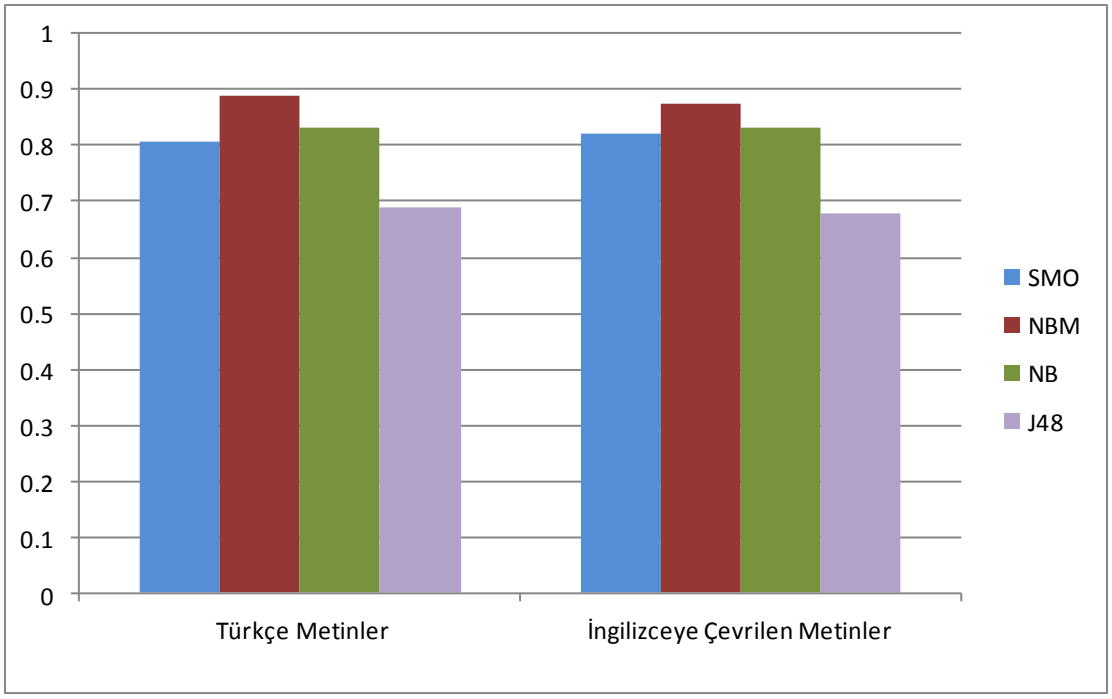
Tablo 11. Türkçeden İngilizceye Çevrilen Metinlerin Sınıflaması (Cross-validation Folds = 10 kullanılarak)

Kategori	SMO	NBM	NB	J48
Ekonomi	0.693	0.826	0.795	0.641
Tıp	0.839	0.895	0.869	0.705
Politika	0.815	0.852	0.805	0.615
Din	0.814	0.85	0.759	0.698
B,ilim	0.842	0.862	0.816	0.644
Spor	0.936	0.964	0.936	0.818
Ortalama	0.822	0.873	0.83	0.677

Tablo 12. Türkçeden İngilizceye Çevrilen Metinlerin Sınıflaması (Percentage split = 66% kullanarak)

Kategori	SMO	NBM	NB	J48
Ekonomi	0.734	0.784	0.779	0.605
Tıp	0.784	0.887	0.882	0.667
Politika	0.735	0.835	0.784	0.62
Din	0.713	0.871	0.788	0.706
Bilim	0.732	0.818	0.792	0.671
Spor	0.871	0.969	0.935	0.798
Ortalama	0.807	0.859	0.825	0.719

Türkçe metinler ve çevirileri üzerindeki kıyaslamalar Şekil 21'de gösterilmiştir.



Şekil 21. Türkçe metinlerin sınıflaması ve İngilizceye çevrilen metinlerin sınıflandırılmalarının kıyaslaması

Türkçe, İngilizce ve çevrilen (Türkçeden İngilizceye ve İngilizceden Türkçeye) metinler sınıflandırılmış, uygulanan yöntem ve kategorilere göre performanslar göre ölçülmüş ve kıyaslanmıştır. Sonuçlar aşağıdaki tablolarda verilmiştir. Trk. Türkçe, İng.

İngilizce, Trk. den İng. Türkçeden İngilizceye ve İng. den Trk. İngilizceden Türkçeye çevrilen metinlerin veri kümesinin temsil etmektedir.

Tablo 13. NB yönteminin sınıflandırma performansı

Kategori	Trk.	Trk. den İng.	İng.	İng. den Trk.
Ekonomi	0.800	0.804	0.930	0.848
Tıp	0.829	0.874	0.941	0.931
Politika	0.793	0.785	0.929	0.878
Din	0.744	0.784	0.947	0.771
Bilim	0.777	0.745	0.914	0.860
Spor	0.904	0.980	0.959	0.940
Ortalama	0.808	0.825	0.935	0.877

Tablo 13'de Naive Bayes algoritmasının performansı kategoriler ilgili veri kümleri üzerindeki ölçüm sonuçlarının vermektedir. Bu sonuçlara göre İngilizce metinlerin sınıflandırılmasında en iyi performansa elde edilmiştir. Diğer yandan, dokümanlar çevrildikten sonra sınıflandırma performansı biraz düşmüştür.

Tablo 14. NBM yönteminin sınıflandırma performansı

Kategori	Trk.	Trk. den İng.	İng.	İng. den Trk.
Ekonomi	0.866	0.774	0.941	0.867
Tıp	0.875	0.874	0.95	0.957
Politika	0.859	0.84	0.934	0.921
Din	0.738	0.849	0.962	0.854
Bilim	0.836	0.826	0.92	0.876
Spor	0.99	0.99	0.97	0.97
Ortalama	0.865	0.859	0.944	0.912

Tablo 14'de Naive Bayes Multinomial yönteminin performansı bütün kategoriler için verilmiştir. Benzer şekilde Naive Bayes algoritması gibi, İngilizce kategorisinde en iyi performans elde edilmiş ve dokümanlar çevrildikten sonra performansın biraz düştüğü görülmüştür.

Tablo 15. SMO yönteminin sınıflandırma performansı

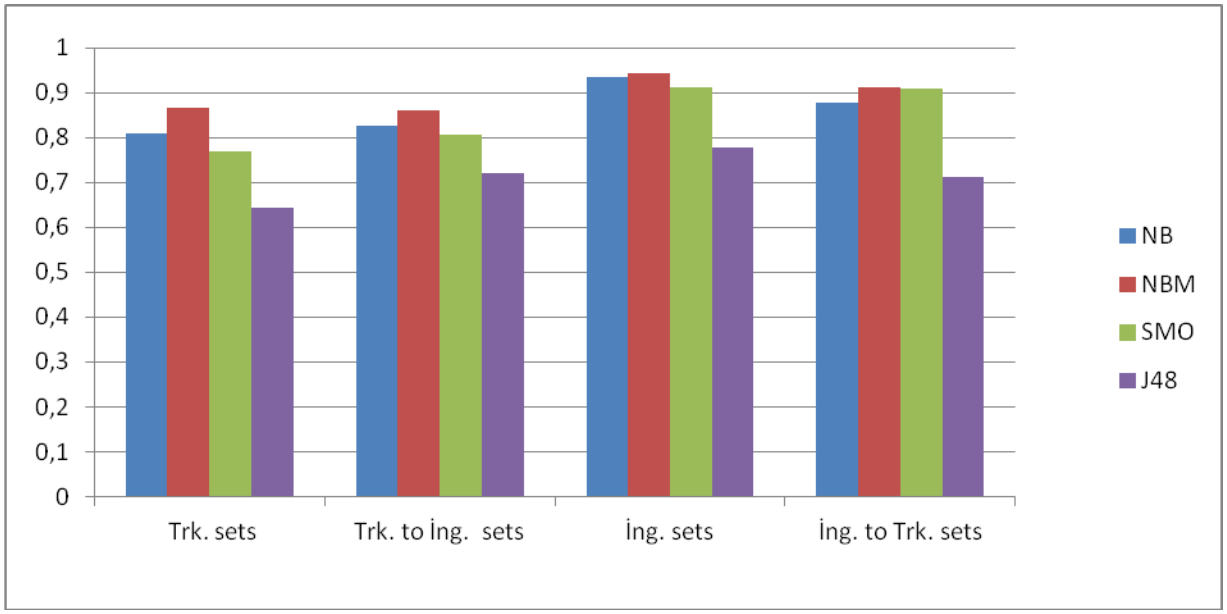
Kategori	Trk.	Trk. den İng.	İng.	İng. den Trk.
Ekonomi	0.76	0.805	0.905	0.851
Tıp	0.744	0.794	0.941	0.966
Politika	0.735	0.779	0.898	0.891
Din	0.677	0.806	0.919	0.889
Bilim	0.782	0.782	0.879	0.898
Spor	0.928	0.898	0.939	0.95
Ortalama	0.77	0.807	0.912	0.909

Tablo 15'de SMO algoritmasının sınıflandırma performansı yine bütün kategoriler için verilmiştir. Yine, İngilizce için en yüksek performans elde edilmiştir ve çevrilen dokümanlar için sınıflandırma performansı düşmüştür.

Tablo 16. J48 algoritmasının sınıflandırma performansı

Kategori	**Trk.**	**Trk. den İng.**	**İng.**	**İng. den Trk.**
Ekonomi	0.66	0.652	0.735	0.519
Tıp	0.627	0.732	0.811	0.842
Politika	0.653	0.643	0.773	0.652
Din	0.522	0.761	0.769	0.737
Bilim	0.631	0.68	0.708	0.695
Spor	0.74	0.904	0.866	0.824
Ortalama	0.644	0.719	0.777	0.711

Tablo 16'da J48 algoritmasının performansı yine bütün kategoriler üzerinde gösterilmiştir. Sonuçlar göstermiştir ki J48 algoritması en düşük sınıflandırma performansını sahiptir. Yine diğer yöntemlerdeki gibi, İngilizcede en yüksek performans elde edilmiş ve çevrinle dokümanlar için performans düşmüştür.



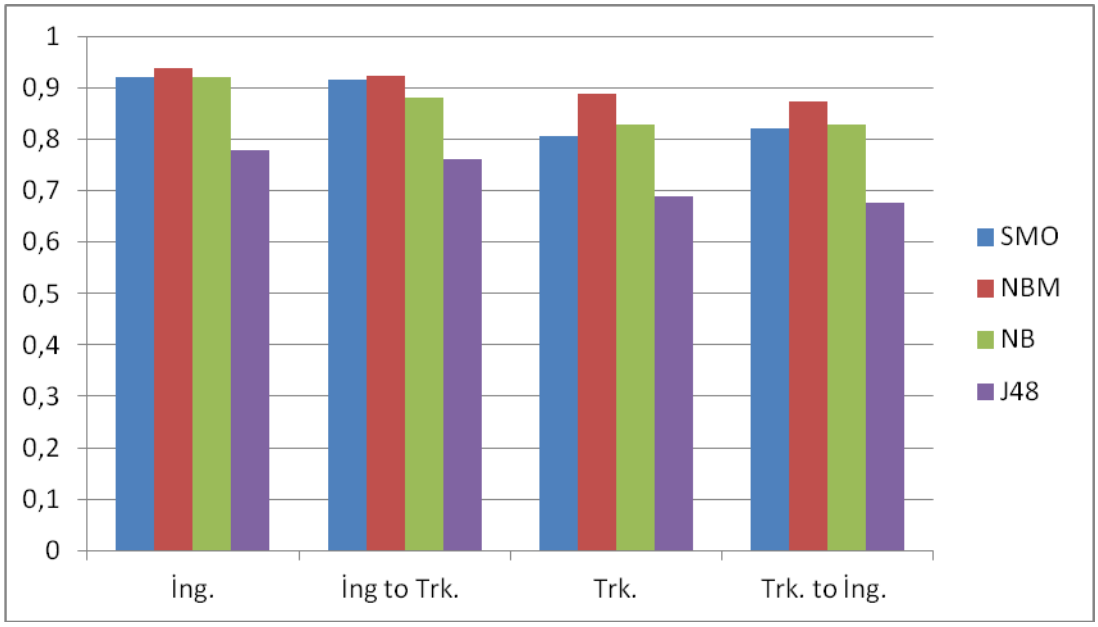
Şekil 22. Algoritmaların performanslarının karşılaştırılması

Şekil 22' de sınıflandırmada kullanılan tüm yöntemlerin ilgili kümeler için karşılaştırılmaları verilmiştir. Sonuçlar göre İngilizcede tüm kategori daha yüksek performansa ulaşılmış ve çeviri yapılan kategorilerde küçük bir yüzdede de olsa bir düşük olduğu görülmüştür.

Tablo 17. Weka ile sınıflama performanslarının karşılaştırılması

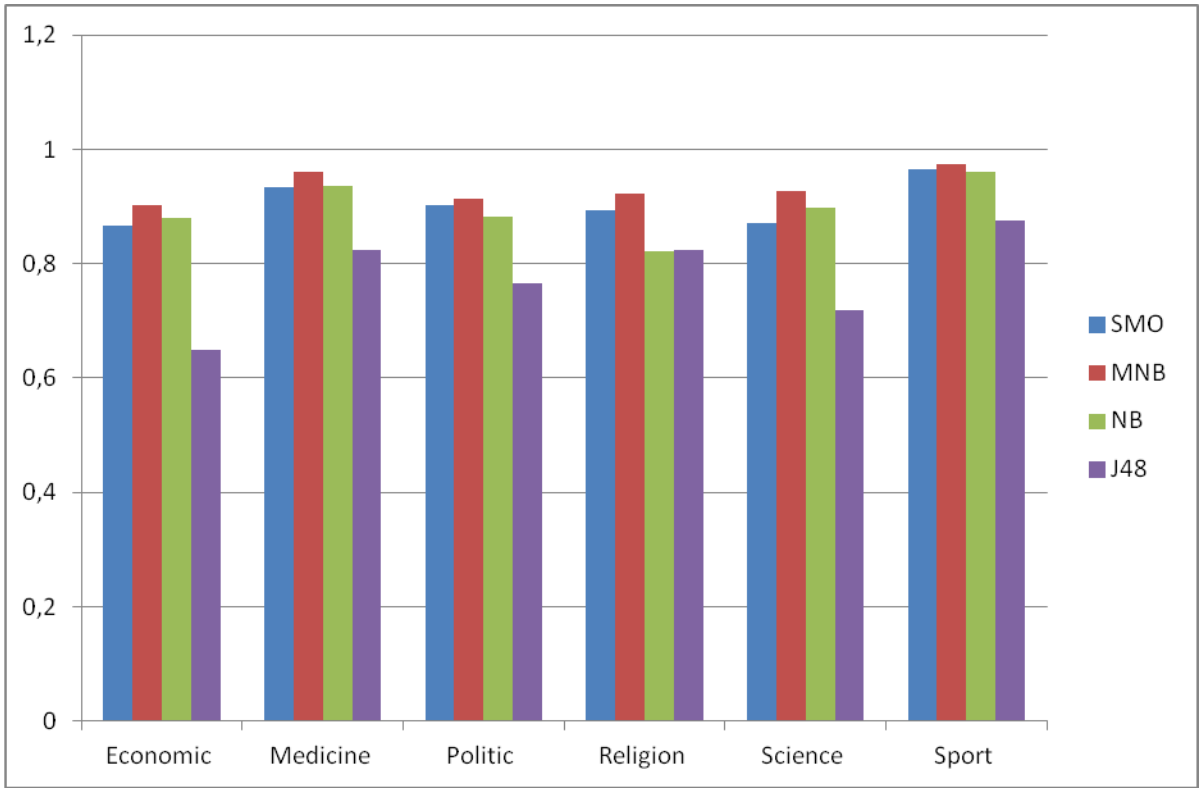
Kategori	**SMO**	**NBM**	**NB**	**J48**
Türkçe	0.807	0.888	0.83	0.69
Türkçeden İngilizceye	0.822	0.873	0.83	0.677
İngilizce	0.921	0.939	0.921	0.78
İngilizceden Türkçeye	0.915	0.923	0.881	0.761

Türkçe, İngilizce metinler ve çeviri metinlerin sınıflandırılma performansları Tablo 17'de verilmiştir. Bu tabloya göre orijinal dilde olan metinlerin sınıflama performansı çeviri metinlere göre daha iyidir. Yine aynı karşılaştırma sonuçları Şekil 23'de farklı bir biçimde verilmiştir.

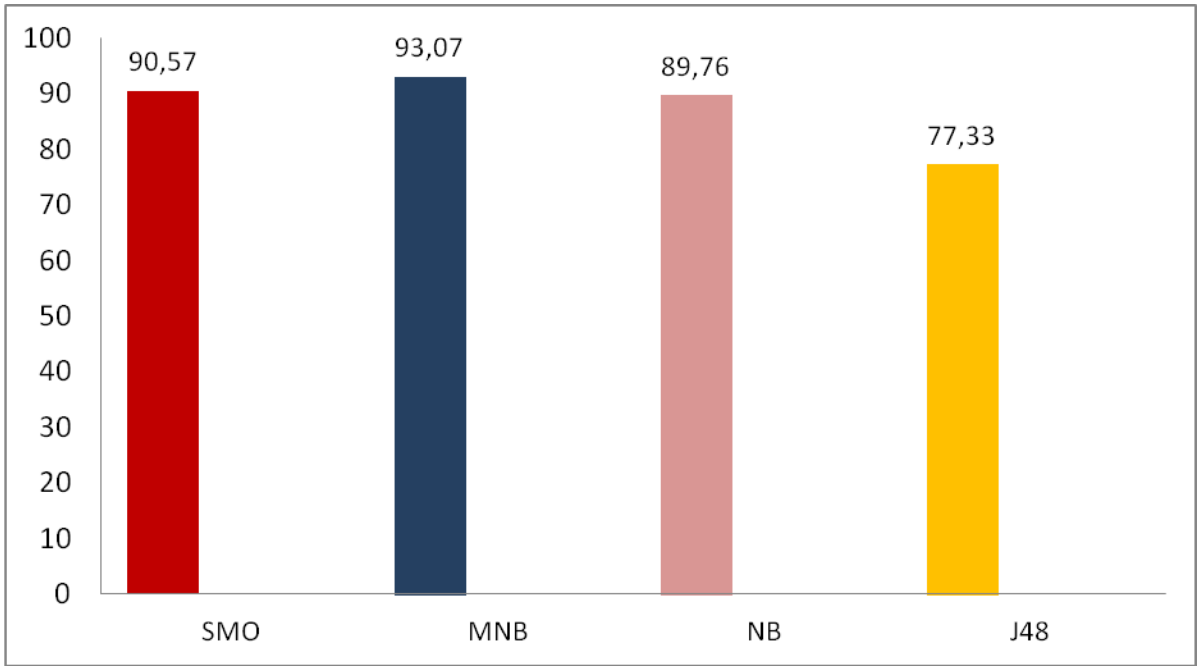


Şekil 23. Weka ile sınıflama performansların karşılaştırılması

Yine aşağıdaki şekillerde (Şekil 24 ve 25) kategorilere ve sınıflandırma yöntemlerine göre sınıflandırma performansları kıyaslamalı olarak verilmiştir.



Şekil 24. Sınıflandırma yöntemlerinin kategorilere göre kıyaslanmaları



Şekil 25. Sınıflandırma yöntemlerinin doğruluk ölçüsüne göre kıyaslanması

4. SONUÇLAR VE ÖNERİLER

Çevrimiçi (online) metinsel bilgilerin büyümesi ile etkin bilgi erişimi, iyi indeksleme olmadan gittikçe daha da zorlaşmaktadır. Metinsel veri sınıflandırma yöntemleri bu soruna oldukça etkili çözümler sunmaktadır. Metinsel doküman sınıflandırma, bir takım belgeleri önceden tanımlanmış kategoriler içinde otomatik olarak sıralama görevdir. Son yıllarda, İngilizce başta olmak üzere, birçok araştırmacı değişik dillerde farlı amaçlara yönelik olarak birçok araştırmalar ve geliştirme çalışmaları yapmaktadır. Bu araştırmaların hemen hemen hepsi orijinal metinler üzerinde yapılmaktadır.

Bu çalışmada, metinsel veri sınıflandırmada bilgisayarlı çeviricilerin etkisi değişik sınıflandırma yöntemleri kullanılarak incelenmiştir. Geliştirilen sistem ilk olarak orijinal dilde yazılan metini analiz edip ve sınıflandırır. Daha sonra aynı metini bilgisayarlı çeviriciler kullanarak analiz yapılacak dile çevirir, aynı yöntemleri kullanarak analiz eder ve sınıflandırır. Dolayısıyla, metinsel veri sınıflandırmada bilgisayarlı çeviricilerin etkisi ölçülmüş, orijinal dildeki sonuçlarla karşılaştırılmış ve değerlendirtmiştir. Yine bu çalışmada, değişik sınıflandırma yöntemlerinin performansları ölçülmüş ve karşılaştırılmıştır. Elde edilen sonuçlara göre Multinomial Naive Bayes sınıflandırıcının en başarılı yöntem olduğu belirlenmiştir. Diğer yandan bilgisayarlı çeviricilerin, aynı metinsel dokümanın farklı dillere tercüme edilmiş olup olmadığına bakılmaksızın, sınıflandırma üzerinde oldukça küçük bir etkisi olduğu tespit edilmiştir. Bu sonuçlar, bilgisayarlı çeviricileri kullanılarak, aynı konularda fakat farklı dillerde de, tek bir dili referans alarak oldukça etkin biçimde veri madenciliği yapılabileceğini göstermektedir.

Yapılan çalışmalarda, bilgisayarlı çevrici olarak yalnızca Google translator kullanılmış, ve orijinal ve çevrilen dokümanların sınıflandırma sonuçlarının değerlendirilmesi sadece Türkçe ve İngilizce için yapılmıştır. Dolayısıyla, gelecek çalışmalar daha fazla dilde daha çok farklı bilgisayarlı çeviriciler kullanılarak, bilgisayarlı çeviricilerin etkileri daha da geniş bir kapsamda ölçülebilir. Yine çevrilmiş dokümanların sınıflandırılma performanslarının orijinal dokümanlara göre azda olsa daha düşük olduğu görülmektedir. Bu durumun başlıca nedenlerinden biri kelimelerin doğru çevrilememesidir. Dolayısıyla, gelecek çalışmalarda uygun disambiguation teknikleri ile sınıflandırma sonuçlarında daha da iyileştirmeler sağlanabileceğini öngörmekteyiz. Diğer yandan,

Türkçe sınıflandırma performansının İngilizce sınıflandırma performansına göre daha düşük olduğu tespit edilmiştir. Burada en önemli nedenler arasında morfolojik analizin uygun biçimde yapılmamış olması gelmektedir. Burada, uygun morfolojik analiz tekniklerinin kullanılması ile sistemin performansı daha da artırılabilir.

5. KAYNAKLAR

1. Adsız A., "Metin Madenciliği", Ahmet Yesevi Üniversitesi, http://www.scribd.com/doc/16574486/31/Boyut-Kucultme, Şubat 2006.

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

3. Baykal A., "Web Madenciliği Teknikleri", Akademik Bilişim'09 Konferansı Şubat 2009, Harran Üniversitesi, Şanlıurfa

4. Bidgoli A. M. and Boraghi M., "A Language Independent Text Segmentation Technique Based on Naive Bayes Classifier", Islamic Azad University North Tehran Branch, M.IEEE; Manchester University England, 2010.

5. Bing L., "Web Data Mining", Exploring Hyperlinks, Contents, and Usage Data, With 177 Figures, Springer-Verlag Berlin Heidelberg 2007.

6. Coşkun C. ve Baykal A., "Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması", Akademik Bilişim'11 Konferansı, Şubat 2011 İnönü Üniversitesi, Malatya

7. Chinchor N., "Evaluation metrics", In Proc.of the Fourth Message Understanding Conference, 22–29, 1992.

8. Dolgun M.Ö., Özdemir T.G. ve Oğuz D., "Veri Madenciliği'nde Yapısal Olmayan Verinin Analizi:Metin ve Web Madenciliği", İstatistikçiler Dergisi 2 (2009), 48-58

9. Guran A., Akyokus S., Bayazit N.G. ve Gurbuz M.Z., "Turkish text categorization using N-Gram words", International Symposium on Innovations in Intelligent Systems and Applications – Yildiz Technical University, Dogus University, Yildiz Technical University , Yildiz Technical University, 369-373, 2009.

10. Gündüz Öğüdücü Ş., "Veri Madenciliği Temel Sınıflandırma Yöntemleri", Ders Notlari, www.cs.itu.edu.tr/~gunduz/courses/verimaden/, 2009.

11. Güngör T., "Lexical and morphological statistics for turkish", International XII.Turkish Symposium on Artificial Intelligence and Neural Networks –, 1-4, 2003.

12. Gorunescu F. “Data Mining”, Concepts, Models and Techniques, Intelligent Systems Reference Library,Volume 12, ISRL 12, 1–43.springerlink.com , © Springer-Verlag Berlin Heidelberg 2011.

13. Kao A., Poteet(Eds) S. R., “Natural Language Processing and Text Mining”, © Springer-Verlag London Limited, 2007.

14. Kavurkaci Ş., Gürkaş Aydın Z., Şamlı R., “Büyük Ölöçekli Veri Tabanlarinda Bilgi Keşfi”, Akademik Bilişim'11 Konferansı, İnönü Üniversitesi, 2-4 Şubat 2011, Malatya.

15. Kesgin F., “Türkçe Metinler için Konu Belirleme Sistemi”, İstanbul Teknik Üniversitesi, 2007.

16. Özdemir S., “ Veri Madenciliği”, Ders notlari, http://w3.gazi.edu.tr/ ~suatozdemir/ , http://ceng.gazi.edu.tr/~ozdemir/

17. Platt J.C., “Fast training of support vector machines using sequential minimal optimization”, MIT Press – Cambridge - MA, 185-208, 1999 .

18. Peng F., Schuurmans D. and Wang S., “ Language and task independent text categorization with simple language models”, Proceeding of HLT – NAACL , School of Computer Science, University of Waterloo - 200 University Avenue West - Waterloo – Ontario - Canada, N2L 3G1, 110-117, 2003.

19. Sak H., Gungor T. ve Saralar M., “ Turkish language resources: morphological parser, morphological disambiguator and Web corpus”, In GoTAL 2008 , vol: 5221 of LNCS , 417-427, 2008.

20. Sasaki Y. and Fellow R., “The truth of the F-measure”, MIB -School of Computer Science, University of Manchester, M1 7DN, 1-5, 2007.

21. Song M. and Brook Wu Y.F., “Text and Web Mining Technologies”, Handbook of Research, Information Science Reference, Hershey. NewYork, Copyright © 2009 by IGI Global.

22. Salton G., A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing,"Communications of the ACM, vol. 18, nr. 11, pages 613–620, 1975.

23. Şentürk A., "Veri Madenciliğin temel Kavram ve Teknikleri", Ekin Basım Yayın, 2006.

24. Torunoglu D., Cakrman E., Can Ganiz M., Akyokus S. ve Gurbuz M.Z., "Analysis of preprocessing methods on classification of turkish texts", IEEE - Department of Computer Engineering Doğuş University-Istanbul-Turkey, 112-117, 2011.

25. Takci H., "Veri Madenciliği", Ders notlari, htakci.sucati.org. http:// htakci.sucati.org /datamining/.

26. Weiss S. M. and Zhang T., "Text Mining", Predictive Methods for Analyzing Unstructured Information, Springer Science+Business Media, Inc., 2005.

27. Yu X., Yuan Y., Tungare M., Perez-Quinones M., Yuan W. and Fox E., "Automatic syllabus classification using support vector machines", IGI Global- USA,Virinia Tech, 61-74,2009.

28. Yang Y. and Pedersen J.O., "A comparative study on feature selection in text categorization", In Proceedings of the Fourteenth International Conference on Machine Learning - San Francisco, 412-420, 1997.

29. http://www.bilyaz.com/index.php/Web-madenciligi.html/

30. http://tr.wikipedia.org/wiki/Veri-madenciliği

31. http://en.wikipedia.org/wiki/Text_mining

32. http://en.wikipedia.org/wiki/Naive_Bayes_classifier

ÖZGEÇMİŞ

Leila ROUKA, 1979 yılında İran'da doğdu. Sırası ile Emirahmadi İlkokulu, Bahoner Ortaokulu ve Yektai Lisesini Iran'da bitirdi. 2002 yılında Azad Islami Üniversitesi Bilgisayar Mühendisliği- Yazılım Bilgisayar Bölümünü bitirdikten sonra, 2009 yılında Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim dalında yüksek lisans yapmaya başladı.