KIRIKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
YÜKSEK LİSANS TEZİ

WEB SAYFALARININ GİZLİ ANLAM ANALİZİ YAKLAŞIMIYLA OTOMATİK OLARAK SINIFLANDIRILMASI

Elvan DUMAN

Ağustos 2013

ÖZET

WEB SAYFALARININ GİZLİ ANLAM ANALİZİ YAKLAŞIMIYLA OTOMATİK OLARAK SINIFLANDIRILMASI

DUMAN, Elvan
Kırıkkale Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi
Danışman: Prof. Dr. Hasan ERBAY
Ağustos 2013, 72 sayfa

Bilgisayar ve ağ teknolojisinin hızlı gelişimi İnternet'in popülaritesini arttırmaktadır. İnternet üzerindeki bilgi miktarının devasa artışı ve web sayfalarının barındırdığı gürültülü bilginin çeşitliliği nedeniyle web sayfalarının içerik sınıflandırması doğal metin sınıflandırmasına göre daha karmaşık ve zordur. Geleneksel bilgi alma metotları dokümanların sınıflandırılabilmesi için terimlerin doküman içerisinde bulunmasını kullanır fakat bunun sonucunda genellikle ilgisiz web sayfaları sonuç olarak döndürülür. Bu çalışmada, web sayfalarını etkili bir şekilde sınıflandırabilmek için Gizli Anlam Analiz temelli otomatik web sayfası sınıflandırma algoritması geliştirilmiştir. Algoritmanın son aşamasında Destek Vektör Makinesi yardımıyla sınıfları birbirinden ayıran eğri çizilmiştir. Ayrıca başarı ve performansı etkileyen terim ağırlıklandırma ve özellik uzayının yüksek boyutluluk problemine çözüm sağlayan özellik seçim yöntemleri üzerinde çalışılmıştır.

Deneysel sonuçlar önerilen sınıflandırma algoritmasının etkinliğini göstermiştir ve dokümanların iyi temsil edildiği bir terim - doküman matrisinin sınıflandırma performansını geliştirdiğini saptamıştır.

Anahtar Kelimeler: Web Madenciliği , Metin Madenciliği, Web İçerik Sınıflandırma, Gizli Anlam Analizi, Destek Vektör Makinesi

ABSTRACT

LATENT SEMANTIC ANALYSIS APPROACH FOR AUTOMATIC CLASSIFICATION OF WEB PAGES CONTENTS

DUMAN, Elvan
Kırıkkale University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering, M.Sc. Thesis
Supervisor: Prof. Dr. Hasan ERBAY
August 2013, 72 pages

The fast development on the computer and network technology has increased the popularity of Web. Due to the gigantic increase in the amount of information on the web and a large variety of noisy information embedded in Web pages, Web page classification is getting more sophisticated and difficult than pure-text classification. Traditional information retrieval methods use terms occurring in document to determine the class of the document, but the retrieve usually results in unrelated web pages. In this study, Latent Semantic Analysis based automatic web page classification algorithm developed in order to effectively classify web pages. The curve separates the document classes plotted by the Support Vector Machine in the final step of the algorithm. We also study on the feature weighting and the feature selection methods which are used to reduce the size of the feature space.

The experimental results demonstrate that the proposed classification algorithm robust and effectively classify the documents, moreover, the results demonstrate that the better the representation of the documents by term - document matrix results in the better classification.

Keywords: Data Mining, Text Mining, Web Content Classification, Latent Semantic Analysis, Support Vector Machine

TEŞEKKÜR

Tezimin hazırlanması aşamasında yardımlarını ve vaktini esirgemeyen, fikir ve düşünceleriyle beni yönlendiren danışman hocam Sayın Prof. Dr. Hasan ERBAY'a, teşekkürlerimi sunarım.

Ayrıca, hoşgörü ve destekleriyle her zaman yanımda olan aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

ŞEKİL Sayfa

ÇİZELGELER DİZİNİ

1. GİRİŞ

Bilgi, insanın varoluşundan bugüne kadar insan hayatında her zaman en önemli unsur olmuştur. İnsanlar yüzyıllar boyunca yeni bilgiler öğrenmeye, bilgiyi kaydetmeye, arşivlemeye ve bilgiye daha hızlı ulaşabilmek için onu özetlemeye ve sınıflamaya ihtiyaç duymuştur. Teknolojinin gelişmesiyle birlikte de insan yaşantısında birçok şey değişime uğramıştır. Bu değişimden belki de en çok bilgi depolama, bilgi arama, bilginin kaydedildiği ortam ve bilgiye erişim şekillerimiz etkilenmiştir. Bilgisayar ve internet teknolojisinin gelişmesiyle birlikte, artık neredeyse yaptığımız tüm işlemler elektronik ortamda saklanmaya başlamıştır. Bir mağazadan yaptığımız alışverişte aldığımız her bir ürün, iade ettiğimiz bir ürün, ödediğimiz miktar, zaman bilgileri ve hatta o mağazaya giriş ve çıkışımızda kamera kayıtları veritabanında saklanmaktadır. Teknolojinin, hızlı ve ekonomik olarak bilginin saklanmasına elvermesiyle birlikte insan yaşamına dair her şey kayıt altına alınmaya başlanmış ve bunun sonucu olarak sayısal ortamda hızla büyüyen veri devasa veri yığınları oluşturmuştur. Bu devası boyutlara varan verinin işlenmesi, anlamlı bilgiler haline dönüştürülmesi, özetlenmesi, sınıflandırılması, gizli ve anlamlı verilerin gün yüzüne çıkarılması önemli bir bilimsel çalışma alanı haline gelmiştir.

Bilgi Çağı, bilişim ve iletişim teknolojilerindeki gelişimin insanlık tarihinde toplumsal, ekonomik ve bilimsel değişimin yönünü yeniden belirlediği ve giderek ağ toplumunun ortaya çıktığı döneme verilen addır [1]. 1980'lerde İnternet'in kullanımının yaygınlaşması ve nihayet 1995'te tamamen serbest bırakılmasından sonra endüstri sonrası terimi yerini enformasyon sözcüğüyle değiştirmiş, kavram Türkçeye Bilişim Çağı ya da Bilgi Çağı olarak yerleşmiştir [2]. Günümüzde "Bilişim Çağı" terimi, 1990'lardan bugüne kadar olan süre için kullanılmaktadır.

Bilgisayar ve internet teknolojisinin gelişmesi ve bunun getirdiği küreselleşmenin sonucu olarak bilginin yönetimi, 1990'lı yıllardan başlayarak büyük bir ivmeyle kuruluş ve ulusların önem verdiği bir konu haline gelmiştir. Microsoft'un

kurucusu Bill Gates'e göre, İnternet, Bilişim Çağı'nın ve 21. yüzyıl'ın ekonomik ve kültürel oluşumunda anahtar rolü oynayacak en önemli teknolojik gelişmedir. Bunun nedeni, internetin veri ve enformasyon alış-verişi için özgün, bağımsız, açık, ölçeklenebilir ve güvenilir bir ortam teşkil etmesidir. Gates'e göre İnternet Çağı'nın en önemli teknolojik uygulaması World Wide Web olmuştur. Web sayesinde İnternet'te veri ve enformasyon aramak, bulmak ve paylaşmak kolaylaşmış, hemen herkesin elinden gelebilen sıradan bir sürece dönüşmüştür [3].

İnternetin gelişimi sayesinde sayısal ortamda üretilen veri miktarı da hızla büyümektedir. Bu hızlı büyüme sonucunda istenilen bilgiye ulaşmak zor bir hal almaya başlamıştır. Bu problemlerin üstesinden gelebilmek için anlamlı bilginin gün yüzüne çıkarılmasıyla ilgilenen veri madenciliği çalışmaları ortaya çıkmıştır. İnternet üzerinde yapılan veri madenciliği çalışmaları da veri madenciliğinin alt dalı olan web madenciliğinin konusudur.

Giriş bölümünün bundan sonraki kısmında metin sınıflandırma ve web içerik sınıflandırma literatürü kısaca özetlenmektedir.

Wen Zhang, Taketoshi Yoshida, Xijin Tang [4], yaptıkları çalışmada doküman gösteriminde genel olarak var olan iki süreç indeksleme ve ağırlıklandırma işlemleri üzerinde durmuşlardır. TF-IDF, LSI ve çoklu-kelime (multi-word) metin gösterimlerini karşılaştırmışlardır. Bu karşılaştırma Çince ve İngilizce dokümanlar üzerinde yapılmış olup deneysel sonuçlar incelendiğinde LSI yönteminin diğer iki yöntemden daha başarılı performans verdiği gözlemlenmiştir.

Chih-Ping Wei ve arkadaşları [5], metin sınıflandırma çalışmalarının tek bir dil ile yazılmış dokümanlar üzerinde olduğuna dikkat çekiyor ve çalışmalarında çok dilli dokümanlar üzerinde sınıflandırma metodu öneriyorlar. Önerilen gizli anlam analizi temelli oluşturulan çok dilli doküman sınıflandırma tekniği tatmin edici doğrulukta sınıflandırma başarısı sağlamışlardır.

Shima K. ve arkadaşları [6], özellik çıkarımında etkili bir metot olan LSI öncesinde özellik sıralama metodu uygulamışlar ve sınıflandırma başarısını artıran

bir çalışma yapmışlardır. Bilindiği gibi LSI özellikleri seçilirken genellikle terim frekansına göz önünde tutulurken sınıflandırma gücü göz ardı edilmektedir. Bu çalışmada destek vektör makinesi temelli bir özellik değerlendirme tekniği önermiştir.

Jiana Meng, Hongfei Lin ve Yuhai Yu [7], özellik seçiminin sınıflandırma başarısındaki rolünün önemli olduğu tespitinde bulunarak metin sınıflandırmasında iki kademeli bir özellik seçim tekniğini çalışması önermişlerdir. Dokümanların vektör uzayında temsili ile çok büyük boyutlu veriler ortaya çıkmakta bu durum sınıflandırma kalitesini ve maliyetini yükseltmektedir. Çalışmada öncelikle terim sayısını düşürmek için novel özellik seçim metodu uygulanmış daha sonra gizil anlam indeksleme temelli yeni anlamsal uzay oluşturulmuştur. Yapılan iki aşamalı özellik seçimi uygulaması diğer uygulamalarla karşılaştırılmış ve daha başarılı sınıflandırma yapıldığı tespit edilmiştir.

Hua Yun ve arkadaşları [8], çeşitli gizli anlam indeksleme sorgu yaklaşımlarını analiz etmiş ve tekil değer yeniden ölçekleme (Singular Value Rescalling, SVR) tekniğini çalışmasında önermişlerdir. Karşılaştırmalar tekrarlamalı artık yeniden ölçekleme (Iterative Residual Rescaling, IRR) ile yapılmış ve başarı oranının %5.9 oranında arttığı gözlemlenmiştir.

Vinay Bhat ve arkadaşları [9], web üzerinden bilgi toplamanın zorluklarına dikkat çekmişlerdir. Bu çalışmada eş ve benzer anlamlı varlıkların tespiti üzerinde durmuşlar, LSA metodunun performansını olumsuz yönde etkileyen yönleri araştırılmış ve gizli anlam analizi temelli iki aşamalı bir algoritma önermişlerdir.

Bo Yu, Zong-ben Xu, Cheng-hua Li [10], yaptıkları çalışmada değiştirilmiş geri yayılımlı yapay sinir ağları ve geri yayılımlı yapay sinir ağları (BPNN) kullanarak yeni bir metin sınıflandırma modeli oluşturmuşlardır. BPNN'in yavaş eğitim hızı sebebiyle değiştirilmiş BPNN algoritması yardımıyla eğitim hızlandırılmıştır. LSA kullanarak ayrıca terim boyutları indirgenmiş ve sınıflandırma başarısı artırılmıştır.

Selma Ayşe Özel [11], otomatik internet sayfalarının sınıflandırılması için genetik algoritma tabanlı bir sınıflandırma çalışması gerçekleştirmiştir. Yapılan çalışmada HTML etiketlerinin terim ağırlıkları farklılaştırılarak sınıflandırma başarımının artırılması amaçlanmıştır. Ayrıca çalışmada başarım Naive Bayes ve K-En Yakın Komşu sınıflandırıcılarıyla karşılaştırması yapılmıştır.

Bu çalışmada, etkili bir web sayfası sınıflandırma sistemi oluşturabilmek için gizli anlam analizi yaklaşımıyla dokümanların ilişkileri ortaya çıkarılmıştır. Elde edilen doküman vektörleri destek vektör makinesi sınıflandırma metoduyla eğitim ve test işlemleri yapılmıştır. Yapılan deneysel gözlemlerde sınıflandırma sisteminin çeşitli özellik çıkarımı ve özellik ağırlıklandırma yöntemleriyle başarı %100'e kadar çıkarılmıştır. Özellikle terim frekans - ters doküman frekansı ağırlıklandırma yöntemi özellik indirgeme süreç basamağında ki-kare testinin sınıflandırma başarısındaki etkisi ortaya çıkarılmıştır.

Tezin geriye kalan kısmı şu şekilde organize edilmiştir.

Tezin ikinci bölümünde, verinin önemi ve veri madenciliği hakkında genel bilgiler verilmiştir. Veri madenciliğinin alt dalı ve tez konusuyla daha çok ilişkili olan web madenciliği ve metin madenciliği konularından daha kapsamlı bahsedilmiştir. Gizli anlam analizi yönteminin matematiksel temeli olan tekil değer ayrışımı yöntemini açıklayıp LSA'nın uygulama basamakları anlatılmıştır. Son olarak da sınıflandırma yöntemi olan destek vektör makinesi hakkında bilgi verilmiştir.

Üçüncü bölümde geliştirilen sınıflandırma algoritmasının amacı ve öneminden bahsedilmiş olup, gerçekleştirilen uygulamanın süreçlerinin ve tercih edilen terim indirgeme, terim ağırlıklandırma, özellik çıkarımı ve sınıflandırma yöntemi hakkında gerekçesiyle birlikte bilgiler verilmiştir.

Dördüncü bölümde gerçekleştirilen sınıflandırma algoritması ile elde edilen sonuçlar ve değerlendirmelere yer verilmiştir. Yapılan uygulamanın sınıflandırma başarısı ve kullanılan yöntemlerin sınıflandırma başarısındaki etkisi deneysel sonuçlarla açıklanmıştır.

Tezin son bölümünde ise yapılan çalışmanın kazanımları ve sonraki çalışmalara katkısı tartışılmıştır.

2. MATERYAL VE METOD

2.1. Bilgi Keşfi

2.1.1. Veri ve Bilgi Kavramı

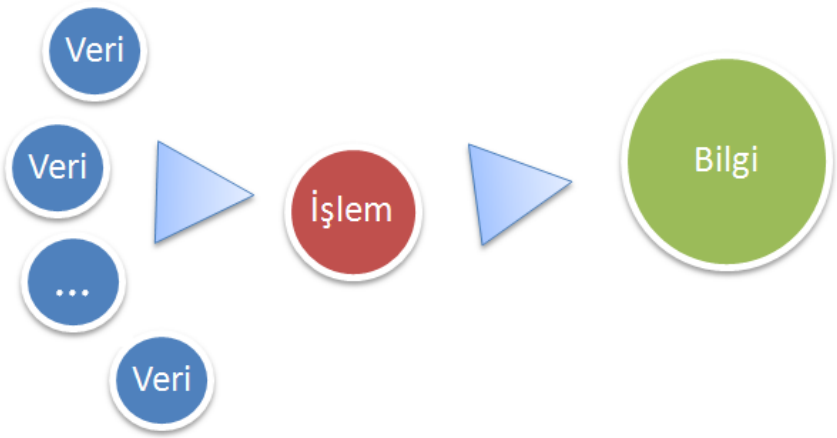
Kavramsal olarak "bir akıl yürütmenin, bir araştırmanın temeli olan ve olduğu gibi kabul edilen öğe" olarak tanımlanan veri; ham, işlenmemiş, kullanılmak üzere olan olay veya durum olabilir [12].

Bilgisayar bilimleri açısından veri, hesaplama ya da manipülasyon amacı ile kullanılan bir gerçeği belirtmektedir. Veriler ya makine düzeyinde ikili gösterimle ifade edilmekte ya da karakterler (harfler veya rakamlar) biçiminde kodlanmaktadır. Her verinin bir türü bulunmakta, bu türlere veri tipleri denmektedir [13].

Veri tek başına değersizdir; ancak bir anlam yüklediğiniz takdirde değer kazanır. İstenilen amaç doğrultusunda bilgi oluşturmak için verinin işlenmesi gerekmektedir. "Bilgi" ise bir soruya yanıt vermek amacıyla veriden çıkarılan sonuç olarak tanımlanabilir. Bu bakımdan veri ve bilgi arasında önemli bir iletişim ağı olduğu söylenebilir. Bilgi anlamlı olarak bazen yalın halde bazen de diğer bilgilerle birlikte çeşitli durumlarda istenilen sonuca ulaşmak için kullanılabilir [12]. Veri ile bilgi arasındaki ilişki Şekil 2.1'de gösterilmiştir.

İnsan yaşantısında bilgi, insanlığın varoluşundan bu yana en önemli yere sahip olmuştur. İnsan doğumundan başlamak üzere hayatını devam ettirebilmek için bilgiye ihtiyaç duymaktadır; çevresinden edindiği gözlemlediği tüm veriyi bilgiye dönüştürmekte ve anlam kazandırmaktadır. Günümüzde artık insan yaşantısındaki verilerin birçoğu sayısal ortama taşınmıştır. Bu veriler bazen insanlar tarafından yorumlanabilirken bazen de insan tarafından yorumlanamayacak boyutlara ulaşır. İnsanoğlu daha çok veriyi kaydedebilmek, veriye hızlı ulaşabilmek ve veriyi çeşitli tekniklerle yorumlayabilmek için verileri sayısal ortama taşımaktadır.

Buradaki ilk büyük zorluk ise verilerin makineler tarafından anlaşılıp yorumlanacağı formatlara dönüştürmesi gerektiğidir. Bu noktada yapılandırılmış ve yapılandırılmamış veri kavramıyla karşı karşıya kalmaktayız.



Şekil 2.1. Veri ile bilgi arasında ilişkinin gösterimi

2.1.2. Yapılandırılmamış Veri ve Büyük Veri Kavramı

Yapılandırılmamış veri, makine tarafından tanınması zor olan belirli bir veri yapısı bulunmayan çeşitli formatlarda bulunan organize edilmemiş enformasyonlardır. Yapılandırılmamış veriler genellikle metin ağırlıklıdır ve bu metinler, tarihler numaralar ve çeşitli olaylar içerebilir. Yapılandırılmamış veriler herhangi bir format ve düzende olmadığı için geleneksel bilgisayar programları tarafından anlaşılması çok zordur. Bu yüzden yapılandırılmış veriler olarak veritabanlarında tutulduğunda ancak bilgisayar sistemleri tarafından anlaşılır olacaktır.

Amerikan Bilgi Teknolojisi Araştırma ve Danışma şirketi olan Gartner, kuruluşların verilerinin %80 'inin yapılandırılmamış veri olduğunu ve 2010-2014 yılları arasında yapılandırılmamış verinin %650 oranında artacağını

bildirilmiştir[14]. Bu da verinin her yıl % 50 oranında artacağı anlamına gelmektedir.

Yapılandırılmamış verinin çok hızlı bir şekilde artması büyük veri (big data) kavramının ortaya çıkmasına neden olmuştur. Bunun sonucunda "Bilgi Çöplüğü" diye tabir ettiğimiz olgu gün yüzüne çıkmıştır. Büyük veri, ilişkisel veri tabanlarında tutulan yapısal verinin dışında kalan, son dönemlere dek çok da kullanılmayan, yapısal olmayan veri yığınıdır.

Yıkılmış olan yaygın bilişimci inanışına göre, yapısal olmayan veri değersizdi, ama büyük veri bize bir şey gösterdi o da günümüzdeki bilgi çöplüğü diye adlandırılan olgudan muazzam derecede önemli, kullanılabilir, yararlı bilgiler çıkarılabilmektedir [15].

Literatürde toplumsal medya paylaşımları, ağ günlükleri, bloglar, fotoğraflar, videolar, log dosyaları gibi değişik kaynaklardan toplanan verilen incelenmesini konu alan birçok bilimsel çalışma yapılmıştır.

Şekil 2.2'de yapılandırılmamış veri örneği gösterilmiştir.

Form Adı:İzin Formu
Fakülte: Mühendislik Fakültesi
Bölüm: Bilgisayar Mühendisliği
Adı Soyadı: Elvan DUMAN
Kullanabileceği izin süresi: Yirmi (20)
Kullanacağı izin süresi: Oniki (12)
İzin Başlama Tarihi:01-08-2013
İzin türü : Yıllık
İzninde bulunulacak yer:İstanbul
Yorum: İzin almasında bir sakınca yoktur.

Şekil 2.2 Yapılandırılmamış Veri, Bir Form Örneği

Şekil 2.2'de verilen örnekte bir izin formu metin dosyası gösterilmiştir. Metin içerikli olan ve oluşturulmasında özel bir gereklilik içermeyen bu formun bilgisayar sistemleri tarafından anlaşılıp yorumlanması oldukça zordur. Bilgi keşfi yapılabilmesi için öncelikle bilgilerin yapılandırılmış hale geçmesi gerekmektedir.

2.1.3. Yapılandırılmış Veri

Yapılandırılmış veri, bilgisayar programları tarafından kolayca anlaşılabilen, belirli bir formatta sunulan iyi organize edilmiş enformasyonlardır. Veri madenciliği uygulamalarında veriler dikkatli bir şekilde hazırlanır. Bilgilerin iki farklı tipte olması beklenir; (a) Sayısal veri (b) kategorik [16]. Sayısal veriler yaş, ağırlık, gelir gibi birbiriyle karşılaştırabilir değerlerdir. Kategorik veriler ise doğru ya da yanlış, sıfır ya da bir gibi ifade edilebilen verilerdir. Eğitim düzeyi, cinsiyet, sigara kullanımı gibi birçok bilgi kategorik veridir.

Çizelge 2.1. Yapılandırılmış Veri Örneği

Kimlik No	**Unvan**	**İzin Kodu**	**İzin Süresi**	**İzin Tarihi**	**Onay Durumu**
123	Arş. Gör	1	12	01-08-2013	1
345	Prof	2	20	08-08-2013	0
678	Doç	2	10	09-08-2013	1
…	…	…	…	…	…

Çizelge 2.1'de yapılandırılmış bir veri örneği gösterilmiştir.

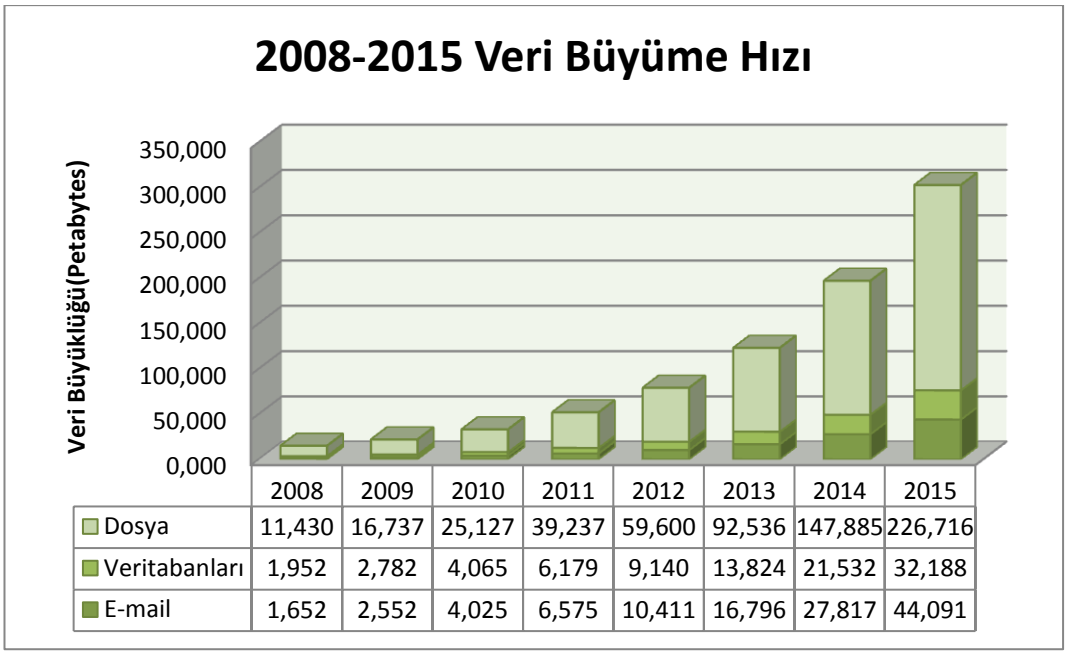
Çizelge 2.1 incelenirse unvan başlığı altında çeşitli akademik unvanlar kategorik olarak sıralanmıştır. Diğer bir kullanım şeklide izin kodu değişkeni gibi bilgilerin alfabetik değil de kodlanmış olarak sunulması yöntemidir. Her bir izin kodu mazeret, hastalık, yıllık gibi farklı izin türlerini ifade etmekte fakat saklanması ve değerlendirilmesindeki kolaylıktan dolayı kodlanmış şekilde veritabanlarında tutulması yoluna gidilmektedir. Metinlerden alınan veriler belli bir sıra ve düzen içerisinde veritabanına kaydedilmiş ve bu veriler artık bilgisayar veritabanı aracılığıyla kolayca ulaşılabilecek istenilen bilgiler değerlendirilip sonuç olarak döndürülebilecektir.

Veri madenciliğinde elektronik tablo halinde sunulan veriler kullanılırken metin madenciliği uygulamalarında metin formatındaki verileri kullanmaktadır. Metin madenciliğinin ana konularından biri metin verilerinin sayısal veri haline dönüştürülüp elektronik tablo şeklinde sunulmasıdır. Böylelikle yapılandırılmamış formatta bulunan veriler, veri madenciliği tekniklerinin uygulanabileceği yapılandırılmış formata dönüştürülebilmektedir [17].

2.2. Veri Madenciliği Kavramı

Verilerden bilgi keşfi olarak da adlandırılan veri madenciliği, büyük boyutlardaki veriden ilginç ve yararlı bilgiyi ve örüntüleri ortaya çıkarma işlemidir [18]. Teknolojinin gelişmesiyle birlikte bilgiyi kaydetmek ve ulaşmak kolaylaşmış bunun aksine bilginin yönetilmesi zor hale gelmiştir. Bilgi ambarları hızla büyümüş ve veri yığınları hızlı bir ivmeyle artmıştır. Bu veri yığınlarından anlamlı ve yararlı bilgileri çıkarma ihtiyacı sonucunda veri madenciliği kavramı ortaya çıkmıştır.

Şekil 2.2'de Yıllara göre çeşitli veri türlerinin artışları ve ileriye dönük tahminleri yer almaktadır [19].



Şekil 2.2. Veri Büyüme Hızı İstatistikleri

Yapay zeka , makine öğrenmesi ve istatistik gibi bilgisayar bilimleri alanındaki bir çok disiplinin veri madenciliği ile bir arada kullanılmaktadır. İstatistik biliminden faydalanarak oluşturulan algoritmalar sayesinde çoklu ve karmaşık analizler oluşturulabilmekte, geleceğe dair farklı senaryolar ortaya konabilmektedir. Böylece veri madenciliği sadece geçmişe dayalı bir istatistiksel çıkarımdan ziyade karar vericiler açısından geleceğe yön vermek amacıyla kullanılmaktadır [20].

Veri madenciliği bankacılık, telekomünikasyon, finans, pazarlama, güvenlik, sigorta, tıp ve birçok alanda uygulanmaktadır. Kar amacı ve Pazar payı elde etmek için pazarlama alanında, müşteri ihtiyaçlarını belirleme ve müşteri profili çıkarabilmek için telekomünikasyonda, hastalık teşhisinde veya hastalığın erken teşhisinde tıp alanında uygulamaları yaygın olarak görülmektedir. Ayrıca, Amerika Birleşik Devletleri'nde gizli dinlemelerde, vergi kaçakçılığının ortaya çıkarılmasında veri madenciliği kullanılmaktadır [21].

2.3. Metin Madenciliği

Metin madenciliği diğer bir adıyla metin veri madenciliği, yapılandırılmamış veya yarı yapılandırılmış veriden yüksek kaliteli bilgilerin çıkarılması işlemidir. Veri madenciliği bilindiği üzere, yapılandırılmış ve kullanılabilir veri üzerinden işlemlerini gerçekleştirmektedir. Metin madenciliği ise yapılandırılmamış verilerden anlamlı en önemli verileri gün yüzüne çıkarma işlemidir. Bu yüzden birbirinin tamamlayıcısı konumunda olduğu için metin madenciliği her zaman veri madenciliğiyle birlikte anılır.

Metin madenciliği uygulamalarının amaçlarından bir tanesi metinden anlamlı ve nitelikli özet bilginin çıkarılmasıdır. Ayrıca binlerce doküman gruplandırılarak belirlenen kategorilere başarılı bir şekilde ayırma çalışmalarını da kapsamaktadır.

Metin madenciliğinde internet ortamındaki verilerin büyük bir önemi vardır, çünkü internet ortamındaki veriler hem iyi yapılandırılmamıştır hem de günden güne hızla büyüyen bir bilgi çöplüğü oluşturmaktadır. Bu sebeple bu devasa bilgilerin içerisinden anlamlı ve değerli bilgilerin çıkarılması büyük bir çalışma alanı yaratmıştır. Fakat bilgi yığınlarının yapısal veri olmaması, büyük bir ivmeyle büyümesi bu alandaki çalışmaların yetersiz kalması ve daha çok araştırma geliştirme çalışmalarının yapılmasını zorunlu kılmıştır.

2.3.1. Metin Madenciliği Kullanım Alanları

Metin madenciliği metinsel içerik bulunan her yerde kullanılabilir. Bilginin artık kağıt üzerinde tutularak ihtiyacı karşılamamaya başladığı yıllardan bu yana metin madenciliği günden güne önem kazanmıştır.
Metin madenciliğinin uygulama alanları arasında şunlar sayılabilir;

- Doküman sınıflandırma,
- Müşteri yorumlarının otomatik analiz edilmesi,
- Metinlerin özetlenmesi,

- Web içeriklerinin sınıflandırılması,
- Soru-cevap sistemleri,
- Metin dili tespiti
- Öneri sistemi,
- Sahtekarlık tespiti,
- Suçluların yazışmalardan anlaşılması,
- Bir metni yazan kişinin tespit edilmesi.

2.4. Web Madenciliği

Web madenciliği, veri madenciliği teknikleri kullanarak internet ortamındaki hizmet ve dokümanlardan değerli bilgilerin çıkarılması işlemdir [22].

Daha önceki kısımlarda sıklıkla günümüzde veriye daha hızlı ve kolay ulaşıldığından ve bu verinin büyük bir ivmeyle arttığından bahsetmiştik. Bu hızlı artışın önemli bir nedeni de bilginin saklanmasında ve kullanılmasında internet teknolojisinin kullanılmasıdır. İnsanlar internet teknolojisi sayesinde bilgiye mekandan bağımsız, hızlı ve ucuz ulaşabilme olanağı bulmuşlardır. İnternet teknolojisinin bu özelliklerinden dolayı internet ortamına herkes tarafından bilgiler saklanmakta ve paylaşılmakta buda bilginin düzensiz bir yapıda yığınlar oluşturmasına neden olmaktadır. Günümüzde World Wide Web üzerindeki internet siteleri ve internet uygulamaları her geçen gün hızla artmaktadır.

Birçok akademisyene göre web madenciliği terimi ilk kez Etzioni tarafından 1996'da "The World Wide Web: quagmire or gold mine" başlıklı makalesinde ortaya atılmıştır. Bu bildiride Etzioni web madenciliğinin, veri madenciliği teknikleri kullanarak World Wide Web'de bulunan dosya ve servislerden otomatik olarak şablon çıkarma, öngörülmeyen gizli ve değerli bilgilerin çıkarılması olduğunu iddia etmiştir [21].

Web madenciliğinin yapabileceği örüntü bulma, sınıflama gibi birçok işlemin arama motorları tarafından yapılabileceği akla gelebilir. Web madenciliğini geleneksel arama motorlarından farklı kılan iki husus vardır. Bunlar:

1. Google, Yahoo, Yandex gibi mevcut arama motorları genel olarak 2 sorunla karşılaşır: Birincisi, bir arama isteği araştırılırken sonuç olarak beklenenden daha çok doküman listelenip kullanıcıya sunulabilir. İkincisi ise getirilen dokümanlar istenilen sorgu için en ilişkiliden en ilişkisize doğru sıralanamamış olabilir. Genellikle bu durumda kullanıcı dokümanlar içinde kendisi arama yapmak zorunda kalır.

2. Geleneksel arama motorları yine bir arama sorgusunda bulunan kelimeler dokümanın içinde geçmediği için yakından ilgili dokümanları sonuç olarak döndürmeye bilir. Örneğin "matematik" konusunun araştırıldığı varsayılırsa bu konu ile yakından ilgili olan *kalkulüs* ve *cebir* ile ilgili dokümanlar sonuç olarak listelenmeyecektir.

Yukarda açıklanan nedenlerden dolayı web madenciliği tekniklerinin arama motorlarıyla birlikte kullanılması döndürülen sonuçların daha tutarlı olmasına katkı sağlayacaktır.

Web madenciliği ile yararlı bilgilerin çıkarılması süreci dört adımda gerçekleşir. Bunlar:

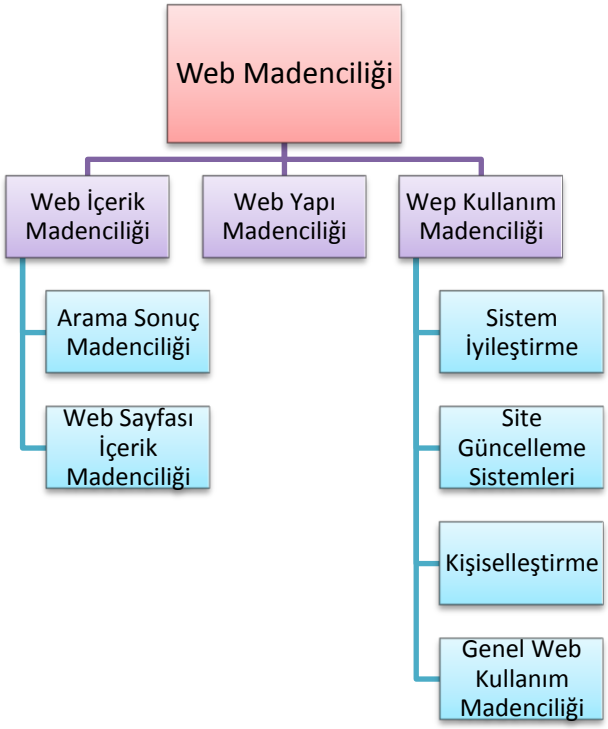
1. **Kaynakların bulunması:** Yapılacak çalışmaya konu olan web sayfası veya web dokümanlarının toplanması.

2. **Bilginin Çıkarılması:** Toplanan web sayfalarının içeriklerinden verilerin çıkarılmasıdır. Çıkarılan veriler henüz yapılandırılmamış ve yararlı verileri çıkarma işlemi henüz yapılmamıştır. Web sayfalarının oluşturulma yapısı ve standartlaşmamış web mimarisi kullanılan sitelerin çokluğu yönüyle metin madenciliğinden ayrılmaktadır.

3. **Genelleştirme:** Veri ön işleme sürecinin ve web sayfalarının arasındaki ilişkilerin tespit edilmesi işlemin yapıldığı aşamadır.

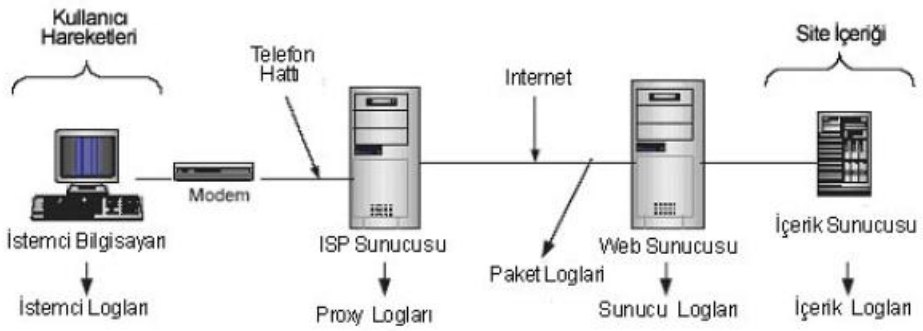
4. **Analiz etme:** Keşfedilen ilişkiler aracılığıyla yorumların çıkarılması sürecidir.

Web madenciliği gelişme yıllarında web içerik madenciliği ve web kullanım madenciliği olarak iki kategoride incelenmekteydi. Son yıllarda web madenciliğinin öneminin artması ve kullanımının yaygınlaşmasıyla yapılan çalışmalar sonucunda üçüncü bir kategori olarak web yapı madenciliği ortaya çıkmıştır. Şekil 2.3'de web madenciliği sınıflandırılması gösterilmiştir.



Şekil 2.3. Web Madenciliği Sınıflandırması

Web madenciliği uygulanmasında veri değişik kaynaklardan toplanmaktadır. Şekil 2.4'de web erişim veri kaynakları ayrıntılı olarak gösterilmiştir.



Şekil 2.4. Web Erişim Diyagramı

Web madenciliğinde kullanılan bu veri tipleri ve özellikleri aşağıdaki gibidir.

Web içerik verisi: İnternet üzerinden kullanıcıların erişebildiği metin, grafik, resim, ses, video gibi belli formatlarda bulunan içeriklerdir.

Web yapı verisi: Bir internet sayfasının organizasyon bilgisidir. Bir web sayfasının kendi sunucusundaki ve diğer internet sayfalarına olan bağlantı bilgileridir. Web sayfasının ağaç yapısı ortaya çıkarılır.

Web kullanım verisi: İnternet kullanıcıların World Wide Web üzerindeki kaynaklara ulaşım zamanı, ulaşım süresi, ulaşım sıklığı gibi bilgi kayıtlarıdır. Web sunucularında bu sunucuya ulaşan istemcilerin IP adresi, ulaşım zamanı, ulaşım süresi, kullanıcının kullandığı tarayıcı tipi ve sürümü gibi birçok bilgiyi log verileri olarak tutmaktadır. Çalışmalar genel olarak bu veriler üzerinden yürütülmektedir.

2.4.1. Web İçerik Madenciliği

Web sayfalarının içeriğinden kullanılabilir yararlı bilgilerin çıkarılma işlemidir. Web sayfalarının içeriği ağırlıkla metinlerden oluştuğu için metin madenciliğiyle

yakından ilgilidir. Web içerik madenciliğini metin madenciliğinden ayıran bazı zorlukları vardır. Web içerikleri genellikle veri biçimindedir, bu özelliğiyle veri madenciliğinden ayrılır. Metin madenciliğinden ayrıldığı nokta ise internet üzerindeki verilerin metin verisi yanında resim ve ses gibi farklı yapılarda bulundurmasıdır.

Web üzerindeki verinin çeşitliliği göz önüne alındığında yararlı bilgilerin çıkarılması çok zor bir görev olarak karşımıza çıkar. Web'de bilginin uçsuz bucaksız olması ve bilginin dinamik olarak değişmesi bilgi kaynağı olarak Web'in kullanılmasını statik veritabanlarının kullanılmasından daha karmaşık hale getirir. Bir diğer önemli nokta ise sorgu sonucunun sunulmasıdır. Devasa büyüklükteki internet verisinden dolayı web sorguları geniş bir sonuç listesi döndürür. Bu yüzden anlamı metotlarla döndürülen geniş listeden en ilişkili içeriklerin kullanıcıya sunulması gerekmektedir [23].

Web içerik madenciliği, kaynaklar arasındaki saklı bilgilerin bulunması ve filtrelenmesini sağlamaktadır. Web kaynaklarından içeriklerine göre otomatik bilgi arama teknikleri tanımlanmaktadır. Bu arama teknikleri iki sınıfa ayrılmaktadır. Bunlar [24];

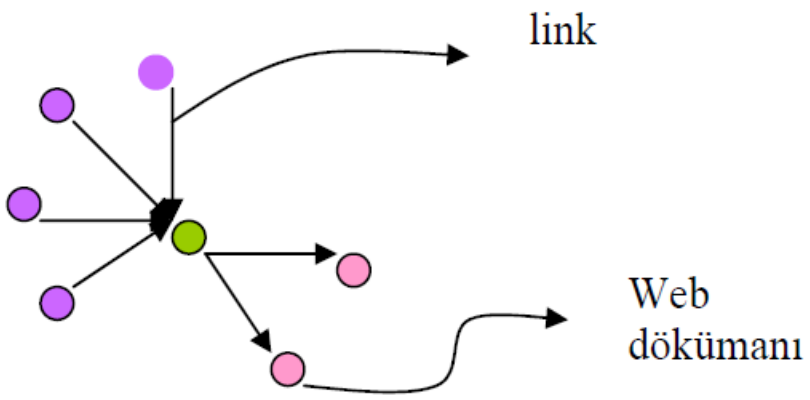
1. **Bilgi Erişim Yaklaşımı:** Kullanıcı profili baz alınarak kullanıcılara gösterilen bilgileri filtrelemek, engellemek ve bilgiye erişimi geliştirmek için kullanılan bir yaklaşımdır.

2. **Veri Tabanı Yaklaşımı:** Web'deki veriyi modellemek ve veriyi bütünleştirerek daha anlamlı bir yapıya sokmak için kullanılan bir yöntemdir. Bu yöntem sayesinde anahtar kelime tabanlı arama yerine daha gelişmiş sorgulama yapabilmek mümkündür.

2.4.2. Web Yapı Madenciliği

Web yapı madenciliği, web sayfasının yapısını bize sunan bağlantı yapılarının incelenmesiyle yararlı bilgilerin ortaya çıkarılmasını sağlar. Bir web sayfasının içeriğine bakılmaksızın, web yapı madenciliği web dokümanları arasındaki bağlantı ilişkisine bakar ve site haritaları çıkarır.

Şekil 2.5'de web grafik yapısı gösterilmiştir. Web dokümanları arasındaki oklar iki sayfa arasındaki ilişkiyi temsil etmektedir.



Şekil 2.5. Web Sayfaları Arasındaki Link İlişkileri

2.4.3. Web Kullanım Madenciliği

Web kullanım madenciliği, web sayfalarının kullanımının araştırılmasıyla yararlı bilgilerin ortaya çıkarılmasıyla ilgilenir. Kullanıcılar tarafından bir web sitesine ulaşım sayısı, her bir ulaşımdaki kullanım süresi, kullanım zamanları, kullanım sırası gibi birçok kullanıcıdan bağımsız davranışların incelenmesini içerir. Bu veriler sunuculardan ve vekil sunuculardan elde edilmektedir. Bu yararlı bilgiler sayesinde site kullanıcı profili çıkarabilir.

Web kullanım madenciliği uygulama alanları Şekil 2.6'daki gibi 5 farklı alana ayrılır [25].

Şekil 2.6. Web Kullanım Madenciliği Uygulama Alanları

Web kullanım bilgileri kuruluşlar için önemli verileri içerir. Kurum bu bilgileri analiz ederek müşteri profilini çıkarıp, müşteri memnuniyetini ölçebilir. Müşterinin ilgi alanlarını, ihtiyaçlarını belirleyerek ürün düzenlemesi ve pazar stratejisi geliştirilebilir.

Çizelge 2.2'de web sayfasını kullanan bir müşterinin sunucu log günlüğü verilmiştir. Log kayıtlarında bağlanan kişinin ip adresi, bağlanma zamanı, bağlandığı adres, veri indirme boyutu, daha sonra ulaştığı sayfalar ve tarayıcı bilgileri gibi değerli veriler yer almaktadır. Bu veriler web madenciliği teknikleri yöntemleriyle analiz edilip kıymetli bilgiler gün yüzüne çıkarılabilecektir. Web kullanım bilgileri sayesinde, web sayfalarının kullanıcı profilini çıkarma, web sayfalarının içeriklerinin uygun olarak güncellenmesi, kişiselleştirme ve sistem iyileştirme gibi bir çok uygulama yapılabilmektedir.

Çizelge 2.2. Örnek Bir Müşteri Server Log Kayıt Örneği

IP adresi	Zaman	Metot / URL / Protokol	Durum	Boyut	Yönlendirme	Agent
87.65.43.21	[25/Nisan/2013:03:04:41]	GET A.html http/1.0	200	2542	-	Mozilla/5.0 (Windows 7)
87.65.43.21	[25/Nisan/2013:03:05:34]	GET B.html http/1.0	200	3140	A.html	Mozilla/5.0 (windows 7)
87.65.43.21	[25/Nisan/2013:03:05:55]	GET A.html http/1.0	200	2542	-	Mozilla/5.0 (windows 7)
87.65.43.21	[25/Nisan/2013:03:06:01]	GET C.html http/1.0	200	8671	B.html	Mozilla/5.0 (windows 7)
87.65.43.21	[25/Nisan/2013:03:08:41]	GET A.html http/1.0	200	2542	-	Mozilla/5.0 (windows 7)
87.65.43.21	[25/Nisan/2013:03:09:21]	GET A.html http/1.0	200	2542	-	Mozilla/5.0 (windows 7)
87.65.43.21	[25/Nisan/2013:03:15:11]	GET C.html http/1.0	200	8671	D.html	Mozilla/5.0 (windows 7)
87.65.43.21	[25/Nisan/2013:03:21:17]	GET D.html http/1.0	200	1680	A.html	Mozilla/5.0 (windows 7)
87.65.43.21	[25/Nisan/2013:03:36:57]	GET E.html http/1.0	200	6172	B.html	Mozilla/5.0 (windows 7)
87.65.43.21	[25/Nisan/2013:04:04:03]	GET B.html http/1.0	200	3140	C.html	Mozilla/5.0 (windows 7)

2.5. Metin Sınıflandırma

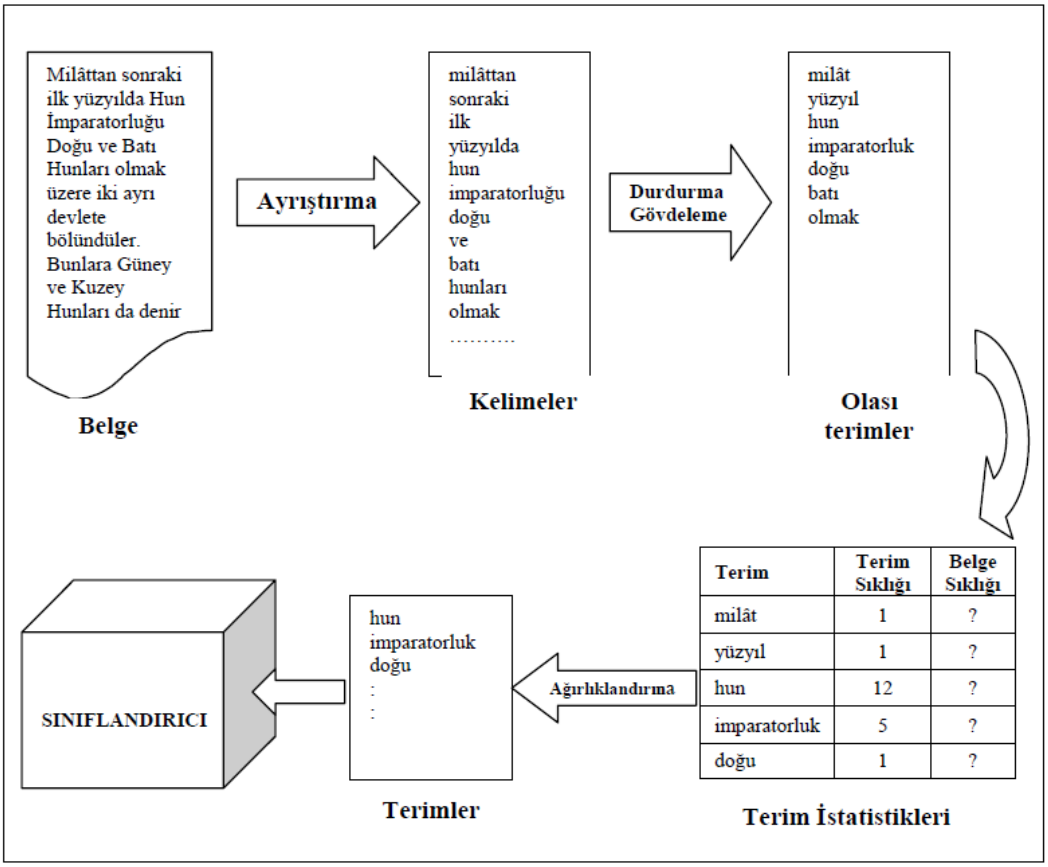
Sınıflandırma, sınıf etiketleri belli bir eğitim kümesi yardımıyla yeni örneklerin ait oldukları kategorileri belirlemeyi amaçlayan bir öğrenme çeşididir. Metin sınıflandırma ise elimizde bulunan mevcut sınıflardan birine ait olduğu bilinen bir dokümanın hangi sınıfa ait olduğunun tespit edilmesi işlemidir.

Günlük yaşantımızda okuduğumuz bir romanın, izlediğimiz bir filmin, dinlediğimiz bir müziğin veya oynadığımız bir oyununun türünü daha önceki bilgilerimizden yararlanarak tespit edebiliriz. Bazen sadece bir sınıfa ait olduğunu tespit ederken bazen de birkaç sınıfa birden ait olduğu veyahut bir sınıfa daha çok yakınken başka bir sınıfa daha az oranda yakın olduğu tespitinde bulunuruz. İşte günlük hayatta yapılabilen bu tespitin bilgisayar sistemleri tarafından otomatik olarak yapılabilmesi amaçlanmaktadır. Bu noktada insanlar tarafından yapılabilen bu tespitin bilgisayar uygulamaları tarafından neden yapılmak istendiği ve sorusu akla gelebilir. Bu soruya cevap olarak şu örnek verilebilir. İnsan kendi

yaşantısında gerekli olan doküman ve bilginin sınıflandırılmasında bile bilgisayar teknolojisi yardımı olmadan zorlanırken bu işin bilgisayar sistemleri tarafından çok daha büyük boyutlardaki bilginin sınıflandırılması, indekslenmesi, filtrelenmesi, otomatik olarak özetlenmesi gibi birçok işlemin daha etkin ve hızlı şekilde yapılabileceği söylenebilir.

Sınıfların oluşabilmesi için öncelikle sınıfı oluşturacak metin verilerinin ortak özelliklerinin çıkarılması gerekmektedir. Ortak özellikleri çeşitli algoritmalarla değerlendirilip sınıf özellikleri meydana getirilir. Test dokümanları sınıf özellikleriyle değerlendirilip çıkan sonucu gerçek sonuçla karşılaştırarak sınıflandırma başarısı ortaya koyulur.

Metin sınıflandırma işlemi gerçekleştirilirken Şekil 2.7'de görülen aşamalar sırası ile uygulanır [26].



Şekil 2.7. Metin Sınıflandırıcı Genel Yapısı

Metin sınıflandırma hakkında genel bilgilerden sonra sınıflandırma türleri hakkında bilgiler verilecektir.

2.5.1. Tekli ve Çoklu Sınıflandırma

Metin sınıflandırma çalışmasında, bir d belgesinin sadece ve sadece bir sınıfa ait olabileceği veya bir sınıfın üyesiyken herhangi bir başka sınıfa da üye olabileceği durumu sınıflamanın tekli veya çoklu sınıflandırma olarak ayrılmasına neden olur. Sadece bir sınıfa ait olunabilen sınıflama türüne tekil-etiketli(single-label) ya da örtüşmeyen sınıflı (non-overlapping categories) denir. Bir belgenin birden fazla sınıfa ait olabileceği sınıflama türüne ise çoklu-etiketli (multi-label) yada örtüşen sınıflı (overlapping categories) adı verilir [27].

2.5.2. Net Sınıflama ve Derecelendirilmiş Sınıflama

Metinleri sınıflandırırken bir d dokümanının bir sınıfa ait olduğu ve bir diğer sınıfa ait olmadığı gibi net kararlar verilebilir. Buna alternatif olarak d dokümanının belli bir sınıfa aitliği derecelendirilerek de ifade edilebilir. Yani bir d dokümanının a sınıfına $\%100$ üye iken b sınıfına $\%60$ ve c sınıfına $\%0$ oranında üye olduğu gibi bir durumdan söz edilebilir.

2.5.3. Metin Kümesi Oluşturma

İnceleme yapılacak dokümanları ifade eden metin kümesi, internet sayfalarındaki içerikler, kütüphanedeki çevirim içi dokümanlar, bir şirketin raporları, kişisel mailler vb. yapılandırılmamış metin içeriği örnek olarak verilebilir. Çalışma yapılacak alandaki veriler toplanarak veri kümesi oluşturulur.

2.5.4. Ayrıştırma

Metin kümesinin kullanılabilir olması için öncelikle metin içerisindeki yararlı bilgilerin ortaya çıkarılması gerekmektedir. Metinlerin kelimelere ayrıştırılması, eşleştirmelerde duyarlılık olabileceğinden tüm harflerin küçük veya büyük harfe çevrilmesi, metinlerde sıklıkla yer alan noktalama işaretlerinin çıkarılması ve tek heceli sözcüklerin arındırılması ayrıştırma adımında yapılan uygulamalardır.

Web içerik madenciliğinde HTML etiketleri ve istemci taraflı kod parçacıklarının temizlendiği ilave bir adıma daha ihtiyaç vardır.

2.5.5. Durdurma Kelimelerinin Çıkarılması

Metin kümesi kelimelerinin çıkarılma işleminden sonra bu kelimelerin hepsini kullanmamız maliyet açısından büyük bir problem yaratmaktadır. Kelimeler içerisinden değerlendirilmede yararlı olmayacak olanların temizlenmesi bu problemin çözümü olarak düşünülür. Bu durumda yapılabilecek ilk uygulama kendi başlarına bir anlamı olmayan tüm metin dosyalarında sıklıkla kullanılan duraklama kelimelerin (ve, sonra, ile, gibi, veya vb.) çıkarılmasıdır. Bu kelimelerinin çıkarılması sonucu kelime sayısını azaltarak işlem kolaylığı sağlanacak hem de yaratacağı ekstra maliyetin önüne geçilmesi sağlanacaktır.

2.5.6. Gövdeleme

Bu aşamada her bir kelimenin eklerinin çıkarılmasıyla kelime kökleri bulunur. Kelimelerin biçimsel benzerlerinin tespit edilmesi anlamına gelmektedir. Böylece, koşucular, koşucu, koşmak, koş ve koşuyorum gibi aynı kök grubundaki kelimeler bir araya getirilerek bu kelimelerin metinde varlığı ortaya çıkarmada her birinin değeri eşit olarak alınır aksi halde farklı kelimelermiş gibi davranacağından özelliklerin ortaya çıkarılamaması ve değerlendirmenin daha zayıf olmasına sebep olacaktır. Bu işlem sonucunda ayrıca terim sayısında azalma sağlanacağından sınıflandırma performansı artar.

Öte yandan kök bulmada karşılaşılabilecek iki sorun vardır. Birincisi, kök bulma işlemi sonucunda farklı anlamları olan kelimeler aynı köke indirilerek yanlış değerlendirme yapılmasına sebep olur. Yani aslında anlamca yakın olmayan terimlere birbirlerinin aynısıymış gibi davranılacaktır. İkinci sorun ise köke ulaşabileceğinden daha az hecenin çıkarılması ile oluşan kökün tam olarak bulunamamasıdır [28]. Kelimeler ile köklerinin tablolarda tutulduğu örnek Çizelge 2.3'de gösterilmiştir.

Çizelge 2.3. Bir Dokümanda Geçen Aynı Köke Sahip Kelimeler Örneği

Kelime	Kök
kazandı	kazan
kazanacak	kazan
kazanılan	kazan
kazanıyor	kazan
kazandılar	kazan
kazanamaz	kazan

Bu yöntem içerik değerlendirmede birçok fayda sağlarken bir takım maliyetlere sebep olmaktadır. Köklerin bulunması, metinlerin binlerce kelime barındırdığı düşünüldüğünde çok fazla işlem zamanı alacak bir süreçtir. Ayrıca kök bulma işlemlerinden sonra kelimeler ve köklerin ilişkisinin veri ambarlarında saklanması gerektiğinden bunun içinde ayrıca yer ayrılması gerekecek ve saklama maliyeti oluşturacaktır.

Gövdeleme işleminin bir kısıtlılığı da inceleme yapılacak dokümanların sadece gövdeleme yapılacak metin dilinde yazılmış olması zorunluluğudur. Her metin dilinin kendine özgü kuralları olduğundan her dil için farklı bir kök bulma algoritması uygulanması gerekecek bu durumda dil bağımsız bir sistem oluşturamama kısıtlılığı doğuracaktır.

2.5.7. Metin Gösterimi

Önceki bölümlerde bahsettiğimiz gibi metin içerikleri yapılandırılmamış veya yarı yapılandırılmış verilerden oluşmaktadır. Makine öğrenmesi, yapay zeka teknikleri ve veri madenciliği algoritmalarının uygulanabilir olması için verilerin yapılandırılmış hale getirilmesi gerekmektedir. Metin sınıflandırma özelinde konuşursak yapısal olmayan verinin yapısal veriye dönüştürülmesinde sıklıkla vektör uzay modeli kullanılır.

2.5.8. Boyut Küçültme

Bir doküman vektörü seçilen tüm terimler için o dokümandaki durumunu belirten sayısal bir değer alır. Eğer terimler dokümanlardan bağımsız olarak seçilirse çok seyrek bir doküman terim matrisi elde edilir. Terim sayısını çok büyük tutmak ve aslında değerlendirmede etkisi olmayacak terimleri değerlendirmeye almak fazla işlem zamanına ve ihtiyaç duyulan çalışma alanı ve belleğin artmasına sebep olur. Hem işlem maliyetini azaltmak hem de değerlendirmede daha iyi sonuçlar elde edebilmek için seçilen doküman değerlendirme kelimelerinde indirgemeye gidilir. Bu durum doküman matrisinin küçülmesine sebep olur.
İndirgeme işleminde iki yaklaşım bulunmaktadır. Bunlar:

1. Özellik Seçimi (feature selection)
2. Özellik Çıkarımı (feature extraction)

Bu yaklaşımlardan detaylı olarak bahsedilecektir.

2.5.9. Özellik Seçimi

Dokümanlar değerlendirme yapılması için gerekli olandan daha fazla terim bulundurur. Terimler arasında kıyaslama yapılarak alt terim kümesi oluşturulur. Amaç dökümanı temsil eden, dokümanın sınıflandırılmasına yardımcı olmayacak kelimelerden arındırılmış en iyi alt kümeyi elde etmektir. Özellik seçiminin getirdiği ekstra maliyeti karşılığında daha iyi sınıflandırma performansı ve

maliyeti elde edilir. Özellik seçimi sonrası maliyet azalır ve sınıflama performansı artar.

En çok bilinen ve sıklıkla kullanılan özellik seçimi algoritmaları arasında belge frekansı (document frequency), bilgi kazanımı (information gain), terim dayanıklılığı (term strength) ve karşılıklı bilgi (mutual information) bulunmaktadır. Yöntemler filtre yöntemleri olarak da bilinirler ve özellikleri benzeşen entropi temelli değer hesabına göre süzerler. Ayrıca, ki kare (chi-square) ve korelasyon katsayısı (correlation coefficient) metotlarının belge frekansından daha iyi sonuçlar verdiği de ispatlanmıştır[29].

2.5.9.1. Ki Kare Özellik Seçim Yöntemi

Bu yöntemde her bir terimin mevcut tüm sınıflara göre ki-kare istatistikleri hesaplanır. Yani her bir terimin doküman sayısı kadar ki-kare istatistik skoru bulunur. Bu istatistik skoru istenilen eşik skorunun altında ise terim alt terim kümesine alınmaz, eğer eşik skorunun üstüne bir değere sahipse alt terim kümesine alınır. Bir terimin ki-kare skorunun yüksek çıkması, bazı sınıflarda diğer sınıflara nazaran daha belirgin bulunması veya diğer sınıflara göre belirgin olarak az bulunmasıyla mümkün olur. Ki-kare skoru yüksek olan bir terim için yüksek olduğu sınıf için belirleyicidir denilir.

N adet belgeden oluşan bir veri kümesinde, c_i sınıfında yer alan *t* terimi için x^2 ki-kare skoru denklem ile hesaplanır [29]. Bu denklemde

$$x^2(t,c_i) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \qquad (0.1)$$

Formülde geçen sembollerin anlamı:

N : toplam doküman sayısı

A : c_i sınıfına ait ve t bulunduran dokümanlar

B : sınıfına ait ve t bulundurmayan dokümanlar

C : sınıfında olmayan fakat t bulunduran dokümanlar

D : ne sınıfında olan ne de t bulunduran dokümanlar

Tüm terimlerin ki-kare skorları hesaplandıktan sonra iki yöntem seçilebilir. Belirli bir ki-kare skorunun üstündekiler terim kümesine alınabilir veya belirli bir terim sayısı seçilerek ki-kare skorlarına göre büyükten küçüğe doğru listelenen terimlerden belirtilen sayı kadar en üstten seçilebilir.

2.5.9.2. Doküman Frekans Özellik Seçim Yöntemi

Bir terimin doküman frekansı, tüm dokümanlar içinden kaç tanesinde bulunduğunun sayısıdır. Yöntemin uygulanmasında tüm terimlerin doküman frekansı bulunur ve frekansı önceden tanımlanmış bir eşik değerinin altında kalanlar değerlendirmede kullanılmazlar.

2.5.10. Özellik (Terim) Ağırlıklandırma Yöntemi

Özellik ağırlandırma yöntemi bir dokümanın terim vektörünü oluştururken her bir terim için kullanılan değerin sadece varlık veya yokluk dışında daha hassas değerler vererek sınıflandırmaya yardımcı olabilmektedir.

Bir dokümanda, bir sözcük bulunuyorsa doküman vektöründe terim için bir, bulunmuyorsa sıfır değeri konulabilir fakat bu durumda hassas bir değerlendirme yapılmış olmaz. Çünkü aynı terim sıklıkla bulunan bir dokümanla sadece bir adet bulunan dokümanla aynıymış gibi değerlendirilecektir. Bu problemin çözümünde ilk akla gelen ise doküman vektöründe terim için sıklık değerinin atanmasıdır. Ağırlık değerinin bulunmasında birçok yöntem kullanılmaktadır. Aşağıda bazı önemli ağırlıklandırma yöntemlerinden bahsedilmektedir.

2.5.10.1. Bit Ağırlıklandırma Yöntemi

Bit ağırlıklandırma yönteminin işlem maliyeti düşüktür fakat hassas olmayan bir ağırlıklandırma yöntemidir. Kelimenin bir dokümanda bulunması durumunda doküman vektöründe kelimenin temsil edildiği terim ağırlığı bir, aksi halde sıfır değerini alır. Yani a_{ij} terimlerinin değeri:

$$a_{ij} = \begin{cases} 1 & f_{ij} > 0 \\ 0 & \text{diğer durumlar} \end{cases} \tag{0.2}$$

ile bulunur. Burada f_{ij} , a_{ij} 'nin frekansını ifade eder.

2.5.10.2.Terim Frekansı Ağırlıklandırma Yöntemi

Terim frekansı (term frequency, TF) yönteminde kelime bir doküman içerisinde geçme sıklığıyla ağırlıklandırılır, yani, kelimenin doküman içindeki frekansı o terimin doküman vektöründeki değeri olarak atanır. Terim frekans ağırlıklandırma yöntemi basit bir yöntem olup işlem zamanı açısından birçok karmaşık ağırlıklandırma yöntemine göre daha düşüktür.

2.5.10.3. Ters Doküman Frekansı Ağırlıklandırma Yöntemi

Ters doküman frekansı (inverse document frequency, IDF) yönteminde her bir terimin dokümanlarda bulunup bulunmadığı araştırılır ve geçtiği doküman sayısı ile orantılı olarak ağırlıklandırılır. Bir terim çok az dokümanda yer alıyorsa bu terimin belirleyiciliği yüksek olduğu söylenir. Öte yandan bir terim veri kümesindeki çoğu dokümanda yer alıyorsa bu terimin sınıflandırma için belirleyiciliğinin düşük olduğu söylenir.

Bir t_i terimi için ters doküman ağırlığı IDF formülü kullanılarak hesaplanır. Tüm dokümanların sayısı N ; t_i teriminin geçtiği doküman sayısı da n_i ile ifade edilirse, t_i 'nin ters doküman ağırlığı $\text{idf}(t_i)$;

$$\text{idf}(t_i) = \log \frac{N}{n_i} \tag{0.3}$$

2.5.10.4. Terim Frekansı –Terim Doküman Frekans Ağırlıklandırma

Terim frekansı – Ters doküman frekansı (terim frequency- inverse document frequency, TF-IDF) yöntemi TF ve IDF yöntemlerinin birlikte kullanılmasıdır. IDF ağırlıklandırılmasının bir eksik yanı terim frekansının göz ardı etmesidir. Bu yöntemde ise TF yönteminin gücü IDF yönteminin gücüyle birleştirilmiş ve daha başarılı sonuçlar veren alternatif bir yöntem elde edilmiştir.

TF-IDF yönteminin özellikleri aşağıda belirtilmiştir [30]:

- Farklı dokümanlarda sık geçmeyen sözcükler sık geçenlere göre daha değerlidir.
- Bir dokümanda bir terimin sık geçmesi seyrek geçmesine göre daha kıymetlidir.
- Dokümanın uzunluğu sözcüğün değerini etkilememektedir.

TF-IDF'in hesaplanma formülü:

$$\text{tf-idf}\left(t,d,D\right) \;=\; \text{tf}\left(t,d\right) \;\times\; \text{idf}\left(t,D\right) \tag{0.4}$$

Burada . t ağırlıklandırılması yapılacak terimi, d veri kümesindeki bir dokümanı ve D veri kümesindeki tüm dokümanların ifade etmektedir.

2.5.11. Özellik Çıkarımı

Özellik çıkarım yöntemleri çok boyutlu bir veri kümesini tanımlamak için gerekli kaynakları basitleştirmek için kullanılır. Özellikle metin kategorizasyon işlemleri göz önüne alındığında, vektör uzay modelinin oluşturduğu yüksek boyutlu ve seyrek yapının gerektirdiği yüksek işlem gücü ve hafıza maliyetini azaltmak özellik yöntemlerinin temel amacıdır [29].

Özellik çıkarım yöntemleri aracılığıyla esasında var olan çok boyutlu özellik uzayı daha düşük boyutlu bir uzaya dönüştürülür.

En bilinen ve çok kullanılan özellik çıkarımı yöntemleri: Temel Bileşen Analizi (*Principal Component Analysis*, PCA) ve Gizli Anlam Analizidir (*Latent Semantic Analysis*, LSA). Günümüzde LSA özellikle metin madenciliği uygulamalarında sıklıkla kullanılan bir yöntemdir [29].

Tez uygulamasında da kullanılan gizli anlam analizi yöntemi ilerleyen bölümlerde ayrıntılı bir şekilde anlatılacaktır. Tezin sonraki kısmında LSA metodunda terim doküman matrisini ayrıştırmada genelde tercih edilen Tekil Değer Ayrışımı (Singular Value Decomposition, SVD) bahsedeceğiz.

2.6. Tekil Değer Ayrışımı (SVD)

1960'lı yıllardan önce sadece lineer sistem analizinde uygulanmış olan matris ayrışımı, son yıllarda yazılım, sınıflandırma, elektronik sinyal filtreleme, matris transformasyonu ve regresyon analizi gibi birçok alanda kullanılmaktadır. Tekil değer ayrışımı en önemli matris ayrışımlarından biridir.

A, m x n boyutlarında bir dikdörtgen matrisi $m \times m$ boyutlu bir ortogonal matris U, $m \times n$ boyutlu bir köşegen matris S ve $n \times n$ boyutlu bir ortogonal matris V 'nin transpozunun çarpımı biçiminde yazılabilir. Yani

$$A = U\,S\,V^T \tag{0.5}$$

burada

$$\begin{aligned} U^T U &= I \\ V^T V &= I \end{aligned} \tag{0.6}$$

eşitliklerini sağlar.

Tekil değer ayrışımı basamaklarını küçük boyutlu bir A matris için yapalım.

Örneğin, 2×3 boyutlu

$$A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$$

matrisini düşünelim. U matrisini hesaplayabilmek için A 'nın transpozu ile başlayalım.

$$A^T = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix}$$

böylece

$$AA^T = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$$

olarak hesaplanır. Şimdi AA^T 'nın öz vektörlerine karşılık gelen öz değerleri bulalım. Bilindiği üzere AA^T 'un öz vektörleri $AA^T \vec{v} = \lambda \vec{v}$ denklemiyle bulunur. Denklemde AA^T yerine koyulduğunda

$$\begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\begin{aligned} 11x_1 + x_2 &= \lambda x_1 \\ x_1 + 11x_2 &= \lambda x_2 \end{aligned}$$

hesaplama yapılırsa

$$\begin{aligned}(11-\lambda)x_1 + x_2 &= 0\\ x_1 + (11-\lambda)x_2 &= 0\end{aligned} \qquad (0.7)$$

denklem sistemini elde ederiz. Bu sistemi λ için çözümlendiğinde

$$\begin{bmatrix}(11-\lambda) & 1\\ 1 & (11-\lambda)\end{bmatrix} = 0$$

$$(11-\lambda)(11-\lambda) - 1.1 = 0 \;\Rightarrow\; (\lambda-10)(\lambda-12) = 0$$

Bu denklemden öz değerler:

$$\lambda_1 = 10, \quad \lambda_2 = 12$$

Bu öz değerlere karşılık gelen öz vektörleri bulmak için bulduğumuz öz değerleri (2.7) denklem sisteminde yerine koyalım.

$\lambda_1 = 10$ için

$$\begin{aligned}(11-10)x_1 + x_2 &= 0\\ x_1 + (11-10)x_2 &= 0\end{aligned}$$

Bu denklem sisteminden

$$x_1 = -x_2$$

buluruz. Bu durumda $x_1 = 1$ aldığımızda $x_2 = -1$'dir. Böylece. $\lambda_1 = 10$ için öz vektörümüz $[1,-1]^T$ olur.

$\lambda_2 = 12$ için

$$\begin{aligned}(11-12)x_1 + x_2 &= 0\\ x_1 + (11-12)x_2 &= 0\end{aligned}$$

Bu denklem sisteminden ise

$$x_1 = x_2$$

ve bu durumda $x_1 = 1$ seçtiğimizde $x_2 = 1$ değerini alır. $\lambda_2 = 12$ için öz vektörümüz $[1,-1]^T$ olur.

Bu özdeğerleri büyükten küçüğe sıralayacak öz vektörlerden oluşan sütun matrisini oluştururuz. Bu örneğimizde $\lambda_1 = 12$'ye karşılık gelen öz vektörü birinci sütunu; $\lambda_2 = 10$'ye karşılık gelen öz vektör ikinci sütunu oluşturur, yani, özvektörler matrisi:

$$\overline{U} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Dikkat edelim ki $\overline{U}$ matrisi ortogonal değildir. Bir sonraki aşama bu matrisi Gram-Schmidt dikleştirme yöntemiyle ortogonal matrise çevirmektir. Önce $\overline{U}$ matrisinin birinci sütun vektörü $\overline{u}_1$'in normalizasyonu ile başlayalım. Bu bize bulmak istediğimiz U matrisinin birinci sütunu u_1'i verecektir. Böylece

$$u_1 = \frac{\overline{u}_1}{\left\|\overline{u}_1\right\|} = \frac{[1,1]^T}{\sqrt{1^2+1^2}} = \frac{[1,1]^T}{\sqrt{2}} = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T$$

Şimdi U matrisinin birinci sütunu u_2'yi bulalım. Burada dikkat etmemiz gereken iki nokta var. Bunlar: (1) u_2 vektörü u_1 vektörüne dik olacak, (2) u_2 vektörünün boyutu bir olacak. Böylece, $\overline{U}$ matrisinin ikinci sütun vektörü $\overline{u}_2$ aldığımızda

$$\begin{aligned} \overline{\overline{u}}_2 &= \overline{u}_2 - u_1 \overline{u}_2 \bullet u_1 \\ &= [1,-1]^T - [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]^T \cdot [1,-1]^T \bullet [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]^T \\ &= [1,-1]^T - [0,0]^T \\ &= [1,-1]^T \end{aligned}$$

ve $\overline{\overline{u}}_2$ vektörü normalize edilirse

$$u_2 = \frac{\overline{\overline{u}}_2}{\left\|\overline{\overline{u}}_2\right\|} [\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}]^T$$

Sonuçta,

$$U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

olarak bulunur. *V*'nin hesaplaması da benzer şekilde $A^T A$ matrisi kullanılarak bulunur.

$A^T A$ matrisine Gram-Schmidt uygulandığında

$$V = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

Öte yandan köşegen matrisin köşegen elemanları özdeğerlerin pozitif karekökü'nün büyükten küçüğe sıralanmasıyla bulunur. Yani bu örnekte

$$S = \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix}$$

Şimdi bulduğumuz matrisleri formülde yerine koyduğumuzda

$$A = U S V^T$$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

2.7. Gizli Anlam Analizi Yöntemi

LSA, dokümanlar ve doküman verisi olan kelimeler arasındaki gizli anlamsal ilişkiyi ortaya çıkartmak için matematiksel bir ayrıştırma yöntemi olan tekil değer ayrışımını kullanan bir metottur. LSA aynı dokümanlarda bulunan kelimelerin benzer anlamlar taşıyacağı prensibine dayanır.

Geleneksel arama yöntemlerinde bir kelimenin varlığı veya yokluğu önemlidir. Aranan kelimenin yokluğu sonuç olarak dokümanın aranan kelime ile ilgili

olmadığı sonucunu verir. Gizli anlam analizi sonucunda ise kelimeler arasındaki korelasyon da ortaya çıkarılır.

Günümüzde internet sitelerinin filtrelenmesi ve engellenmesi önemli bir çalışma konusudur. İnternet sayfalarının belirlenen sınıfa girip girmediğinin tespit edilmesinde çeşitli yöntemler kullanılmaktadır. Örnek vermek gerekirse, oyun sitelerinin engellenmesini düşünelim. Geleneksel yöntemlerin bazıları doğrudan ilgili kelimelerin varlığı ile tespit yapmaktadır. Bu durum bazen bir arama kelimesinin varlığı yüzünden gerçekte aranan içerik olmamasına rağmen engelleme yapılmasına ya da tam tersi olarak bu durumun farkında olan internet sitesi sahiplerinin engellemeye takılmamak için oyun kelimesi yerine atari gibi eş anlamlı ya da yakın anlama gelebilecek kelimeleri kullanarak sisteme yakalanmamasına neden olur. Bilgisayar oyunları ile ilgili internet sayfalarına gizli anlam analiz uyguladığında, bu internet sayfalarında sıklıkla oyun, atari, beceri, macera kelimeleri geçiyorsa bu kelimeler bu kategori için birbiriyle ilişkili olduğu sonucuna varılır ve değerlendirmede etkili olurlar.
Gizli anlam analizin bir diğer avantajı da yapay sinir ağları ve diğer birçok sınıflandırma yöntemlerinin aksine gözetimsiz öğrenme yöntemidir. Gözetimsiz öğrenme yöntemlerinde eğitim setinin bulunması ve karar sürecine geçmeden önce eğitim yapılması işlemlerine gerek duyulmaz ayrıca gözetimli öğrenme yöntemlerinde kullanılacak sınıflar önceden belli olma kısıtlılığı gizli anlamsal analiz için geçerli değildir.

LSA yöntemi hem kelimeler arası ilişki hem de dokümanlar arası ilişkiyi betimleme özelliğine sahiptir. Yani bir kavram uzayı, birbiriyle ilişkili terimler ve bu terimlerden oluşan dokümanlar arası ilişki tekil değer ayrışımı yöntemiyle oluşturulur. Bu ilişki kurulduğunda kavram uzayı üzerinde hem dokümanlar hem de terimlerin konumları hesaplanır. Bu çalışmanın özelinde internet sayfaları bu kavram uzayı üzerinde konumlandırılır. Bu konumlandırmada birbiriyle ilişkili olan sayfalar benzerlikleri oranında birbirine yaklaşacak ve böylece sınıflama imkanı sağlayacaktır.

Bir terim-doküman matrisinde sütun uzayı, dokümanları ifade eder ve kavram uzayında dokümanların konumlandırılmasını sağlar. Satır uzayı ise her bir terimi

yani kelimeyi ifade eder ve kavram uzayında terimlerin konumlandırılmasını sağlar.

Bir terim-doküman havuzumuzda, *d* doküman sayısını ve *t* terim sayısını belirtilir ve bir A matrisi ile gösterilirse bu matris U, S ve V matrislerinden oluşan faktörlerine ayrılır:

$$A = U S V^T \tag{0.8}$$

$U : t \times t$ ortogonal bir matris olup sütunlarında *A*'nın sol tekil vektörlerini taşır.

$V : d \times d$ ortogonal bir matris olup sütunlarında *A*'nın sağ tekil vektörlerini taşır.

$S : t \times d$ köşegen matris olup, köşegen elemanları *A*'nın tekil değerleridir.

S matrisinin elemanları

$$\sigma_1 \geq \sigma_2 \geq \sigma_1 \geq \sigma_3 \geq \sigma_4 \geq \cdots \geq \sigma_{\min(t,d)} \geq 0 \tag{0.9}$$

biçimindedir.

Şimdi bir örnek üzerinde LSA'nın basamaklarını uygulayalım.

Örnek : 6 adet konu başlığından oluşan doküman başlıkları ve seçilen terimler aşağıdaki gibi olsun.

Dokümanlar:

d1: Ünlü oyuncunun gösterisi bu ay sahneye çıkıyor.
d2: Filmin galasına tüm yönetmen ve oyuncular katıldı.
d3: Yazar son eseriyle en çok satan kitaplar listesinde ilk sırada
d4: Film yazarın kitaptaki açısından çok uzak.
d5: Filmin son sahnesinin çalıntı olduğu iddia edildi.
d6: Yazar gazetelere kitap hakkında röportaj verdi.

Terimler:

T1 : ünlü

T2 : oyuncu

T3 : sahne

T4 : film

T5 : gösteri

T6 : yazar

T7 : kitap

T8 : eser

Terim-doküman matrisi, $t \times d$ boyutludur ve

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

A'nın tekil değer ayrışımı hesaplandığında

$$U = \begin{bmatrix} -0,0488 & 0,3536 & 0,3701 & 0,0000 & -0,1654 & -0,4564 & -0,1308 & 0,6949 \\ -0,1206 & 0,5238 & 0,0756 & 0,7071 & 0,1152 & 0,4382 & 0,0000 & 0,0000 \\ -0,1206 & 0,5238 & 0,0756 & -0,7071 & 0,1152 & 0,4382 & 0,0000 & 0,0000 \\ -0,3779 & 0,3213 & -0,7849 & 0,0000 & 0,0551 & -0,3672 & 0,0000 & 0,0000 \\ -0,0488 & 0,3536 & 0,3701 & 0,0000 & -0,1654 & -0,4564 & 0,1308 & -0,6949 \\ -0,6242 & -0,2084 & 0,1566 & 0,0000 & -0,1812 & 0,0980 & -0,6949 & -0,1308 \\ -0,6242 & -0,2084 & 0,1566 & 0,0000 & -0,1812 & 0,0980 & 0,6949 & 0,1308 \\ -0,2101 & -0,1052 & 0,2224 & 0,0000 & 0,9220 & -0,2128 & 0,0000 & 0,0000 \end{bmatrix}$$

$$S=\begin{bmatrix} 2,6348 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2,2278 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1,5519 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0,7791 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0,2818 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

ve

$$V=\begin{bmatrix} -0,1286 & 0,7877 & 0,5744 & 0 & -0,1289 & -0,1286 \\ -0,1892 & 0,3794 & -0,4571 & 0,7071 & 0,2186 & 0,2521 \\ -0,5535 & -0,2343 & 0,3452 & 0 & 0,7183 & -0,0600 \\ -0,6172 & -0,0429 & -0,3039 & 0 & -0,3943 & -0,6077 \\ -0,1892 & 0,3794 & -0,4571 & -0,7071 & 0,2186 & 0,2521 \\ -0,4738 & -0,1871 & 0,2018 & 0 & -0,4651 & 0,6953 \end{bmatrix}$$

bulunur. Öte yandan,

$$V^T=\begin{bmatrix} -0,1286 & -0,1892 & -0,5535 & -0,6172 & -0,1892 & -0,4738 \\ 0,7877 & 0,3794 & -0,2343 & -0,0429 & 0,3794 & -0,1871 \\ 0,5744 & -0,4571 & 0,3452 & -0,3039 & -0,4571 & 0,2018 \\ 0 & 0,7071 & 0 & 0 & -0,7071 & 0 \\ -0,1289 & 0,2186 & 0,7183 & -0,3943 & 0,2186 & -0,4651 \\ -0,1286 & 0,2521 & -0,0600 & -0,6077 & 0,2521 & 0,6953 \end{bmatrix}$$

S matrisinin sondan iki satırı sıfır olduğu için A matrisinin rank'ı altıdır. Bu da sınıflandırma sistemi için altı temel kavramın önemli olması demektir. Bu kavramlar A'nın tekil değerleri olan S matrisinin köşegen elemanları 2.6348, 2.2278, 1,5519, 1, 0.7791 ve 0,2818'dir ve azalan önem dereceleriyle sıralanmıştır. Bu tekil değerlerden küçük olanlar gizli anlam analiz yönteminde göz ardı edilebilir. Bu tekil değerlere bakılarak S matrisi için k tane önemli kavram seçilir ve diğer tekil değerler için sıfır yazılır. Böylece S matrisi S_k matrisine indirgenmiş olur. Ayrıca U matrisi U_k'ya ve V^T matrisi V_k^T matrisine indirgenmiş olur. Bu durumda A matrisi yaklaşık olarak

$$A_k = U_k S_k V_k^T \tag{0.10}$$

matrisine eşittir.

Yukarıdaki doküman örneğimizde k değerini 2 seçtiğimizde maliyeti azaltılmış ve yaklaşık değeri aynı olan bir doküman terim sistemi elde ederiz. Bu durumda

$$U \approx U_2 = \begin{bmatrix} -0,0488 & 0,3536 \\ -0,1206 & 0,5238 \\ -0,1206 & 0,5238 \\ -0,3779 & 0,3213 \\ -0,0488 & 0,3536 \\ -0,6242 & -0,2084 \\ -0,6242 & -0,2084 \\ -0,2101 & -0,1052 \end{bmatrix}$$

$$S \approx S_2 = \begin{bmatrix} 2,6348 & 0 \\ 0 & 2,2278 \end{bmatrix}$$

$$V \approx V_2 = \begin{bmatrix} -0,1286 & 0,7877 \\ -0,1892 & 0,3794 \\ -0,5535 & -0,2343 \\ -0,6172 & -0,0429 \\ -0,1892 & 0,3794 \\ -0,4738 & -0,1871 \end{bmatrix}$$

$$V^T \approx V^T{}_2 = \begin{bmatrix} -0,1286 & -0,1892 & -0,5535 & -0,6172 & -0,1892 & -0,4738 \\ 0,7877 & 0,3794 & -0,2343 & -0,0429 & 0,3794 & -0,1871 \end{bmatrix}$$

Dikkat edelim ki S_k, U_k ve V_k matrislerinin boyutları sırasıyla $t \times k$, $k \times k$ ve $k \times d$ 'dir. Bu gizli kavramlar içerisindeki terimlerimiz $t \times k$ boyutlu matris

$$U_k S_k \tag{0.11}$$

satırlarıdır. Dokümanlar ise $t \times k$ matris

$$S_k V_k^T \tag{0.12}$$

Sütunlarıdır. Yukarıdaki önerimize geri döndüğümüzde

$$U_2 S_2 = \begin{bmatrix} -0,129 & 0,788 \\ -0,318 & 1,167 \\ -0,318 & 1,167 \\ -0,996 & 0,716 \\ -0,129 & 0,788 \\ -1,645 & -0,464 \\ -1,645 & -0,464 \\ -0,554 & -0,234 \end{bmatrix}$$

$$S_2 V_2^T = \begin{bmatrix} -0,339 & -0,499 & -1,458 & -1,626 & -0,499 & -1,248 \\ 1,755 & 0,845 & -0,522 & -0,096 & 0,845 & -0,417 \end{bmatrix}$$

olduğundan örnekteki terimleri ve dokümanları temsil eden terim ve doküman vektörleri şunlardır:

$$\text{ünlü}=\begin{bmatrix} -0.129 \\ 0.788 \end{bmatrix},\ \text{oyuncu}=\begin{bmatrix} -0.318 \\ 1.167 \end{bmatrix},\ \text{sahne}=\begin{bmatrix} -0.318 \\ 1.167 \end{bmatrix},\ \text{film}=\begin{bmatrix} -0.996 \\ 0.716 \end{bmatrix}$$

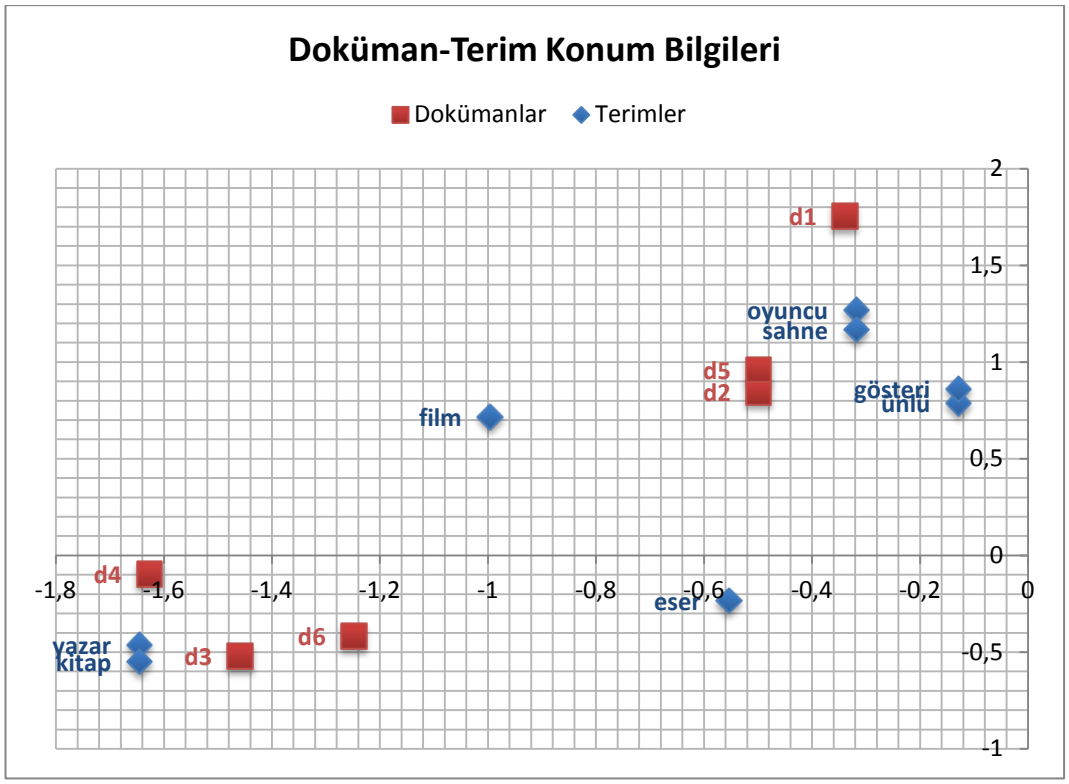
$$\text{gösteri}=\begin{bmatrix} -0.129 \\ 0.788 \end{bmatrix},\ \text{yazar}=\begin{bmatrix} -0.165 \\ -0.464 \end{bmatrix},\ \text{kitap}=\begin{bmatrix} -0.165 \\ -0.464 \end{bmatrix},\ \text{eser}=\begin{bmatrix} -0,554 \\ -0.234 \end{bmatrix}$$

ve

$$\text{d1}=\begin{bmatrix} -0.339 \\ 1.755 \end{bmatrix},\ \text{d2}=\begin{bmatrix} -0.499 \\ 0.845 \end{bmatrix},\ \text{d3}=\begin{bmatrix} -1.458 \\ 0.522 \end{bmatrix},\ \text{d4}=\begin{bmatrix} 1.626 \\ 0.096 \end{bmatrix},\ \text{d5}=\begin{bmatrix} -0.499 \\ 0.845 \end{bmatrix},$$

$$\text{d6}=\begin{bmatrix} -1.248 \\ -0.417 \end{bmatrix}$$

Şekil 2.8'de doküman ve konum vektörleri elde edildikten sonra bu bilgilerle oluşturulan konum grafiği verilmiştir.



Şekil 2.8. Örnek Terim-Doküman Grafiği

Şekil 2.8 incelendiğinde dokümanların ve terimlerin birbirine olan yakınlığı veya uzaklığından anlamca veya içerik yönünden birbirlerine uzak veya yakın yorumu yapılabilir. Örnekte d2 ve d5 dokümanlarının birbirine yakınlığı gözükmektedir. Bu yakınlık bu dokümanların içerik yönünden birbirine yakın olduğunu göstermektedir. Aynı şekilde yazar ve kitap terimlerinin de birbirine olan yakınlığından dolayı bu terimlerin birbiriyle ilişkili terimler olduğunu söylenebilir. Dokümanlardan d1 ve d3 dokümanlarının birbirine çok uzak konumlandığından bu dokümanlarının içeriklerinin birbiriyle uzak ilişkili olduğu anlamı çıkarılır. Benzer şekilde kitap terimiyle oyuncu terimi birbiriyle uzak ilişkilidir denilebilir.

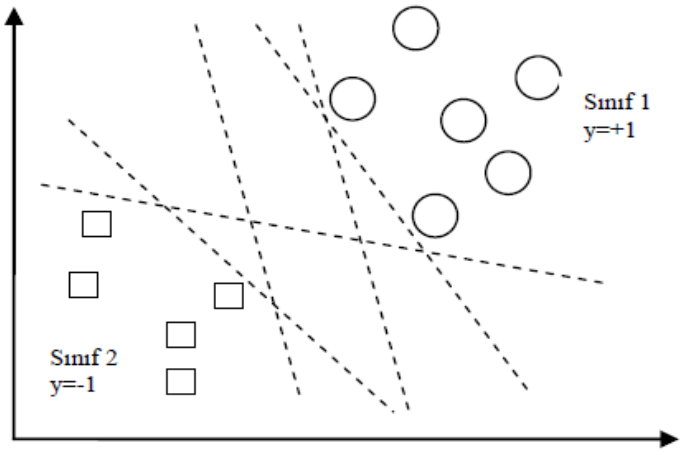
2.8. Destek Vektör Makinesi

Destek vektör makinesi (*Support Vector Machine*, SVM), sınıflandırma problemlerinde kullanılan makine öğrenmesi yöntemleri arasında yer alan bir yöntemdir.

Daha çok makine öğrenmesiyle yöntem, sınıflandırmayı bir doğrusal veya doğrusal olmayan bir fonksiyon yardımıyla yerine getirir. Destek vektör makinesi yöntemi veriyi birbirinden ayırmak için en uygun fonksiyonun tahmin edilmesi esasına dayanır.

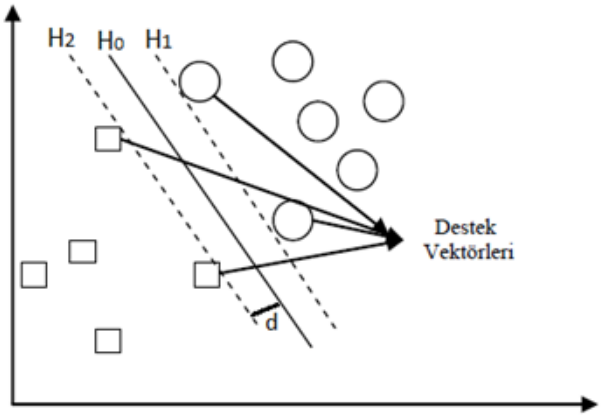
2.8.1. Verilerin Doğrusal Olarak Ayrılabilme Durumu

D, n elemanlı bir veri kümesi ve elemanlarının $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ olduğunu varsayalım. (Şekil 2.9) D kümesinin grafiği iki sınıf halinde versin. Grafikten D kümesinin elemanlarının birbirinden farklı biçimlerde doğrusal olarak ayrılabileceği görülmektedir. Eğer veri kümesini çok boyutlu uzaydan seçersek doğrunun yerini hiper düzlem alacaktır. Fakat, anlaşılması daha kolay olduğundan iki boyutlu düzlemi düşüneceğiz.



Şekil 2.9. Doğrusal Olarak Ayrılabilen Verilerin Görünümü

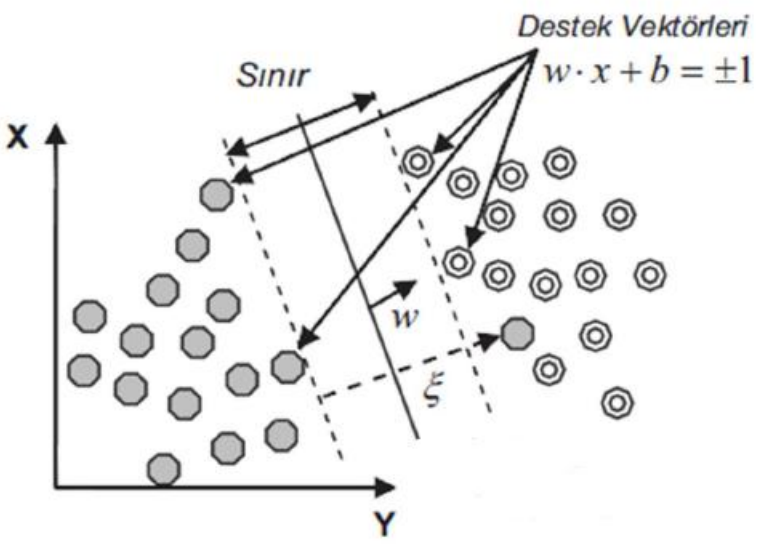
Şimdi amacımız düzlemdeki verileri birbirinden ayıran bu doğrulardan hangisinin veri kümesi sınıflarının her ikisine de en uzak olanını bulmaktır. Bunun için görsel yargımıza dayanarak Şekil 2.10 üzerinde yer alan H_1 ve H_2 doğrularını düşünelim. Bu iki doğrunun ortasından geçen H_0 doğrusu bu iki sınıf veriyi birbirinden en iyi ayıran doğrudur. Bunun teorik nedenleri için "*Web Data Mining*, Exploring Hyperlinks, Contents and Usage Data [31]" isimli kitabın 109-117 sayfalarına bakınız.



Şekil 2.10. Veriler Arasındaki Muhtemel En Büyük Boşluk

2.8.2. Verilerin Doğrusal Olarak Ayrılamama Durumu

Yukarıdaki veri kümesi elemanları öyle iki sınıfa ayrılabilir ki herhangi bir doğru tarafından birbirinden ayrılamaz (Şekil 2.11). Bu durumda sınıfları ayırmada negatif olmayan ve hataları temsil eden gevşek değişken anılan değerler vasıtasıyla sınıflar birbirinden sözde doğru anılan eğriler tarafından ayırmaya çalışılır. Bu eğriler doğrusal olmayan sınıflandırıcılar anılırlar.



Şekil 2.11. Birbirinden Doğrusal Olarak Ayrılamayan Veriler

Gerçek hayatta karşılaştığımız problemler genelde doğrular tarafından ayrılamayan sınıflara ayrılırlar.

3. WEB SAYFALARI SINIFLANDIRMA ALGORİTMASI

Daha önceden bahsettiğimiz gibi bu tez çalışmasında günümüzde hızla sayısı artan web sayfalarının hızlı ve otomatik şekilde sınıflandırılmasının gerçekleştirmesi amaçlanmıştır. Bunun yanı sıra metin dilinden bağımsız sınıflama algoritması geliştirmek bu çalışmanın temel hedefidir. Geliştirilen algoritmaya sınıflandırmada kullanılacak sınıf sayısının değişken olması yeteneklerinin kazandırılması sağlanmıştır. Literatürde web içerik madenciliği çalışmalarında yapılan teorik veya uygulama ağırlıklı çalışmaları internet ortamına taşımak ve çevrimiçi gerçekleştirmek çalışmanın uygulama hedeflerindendir.

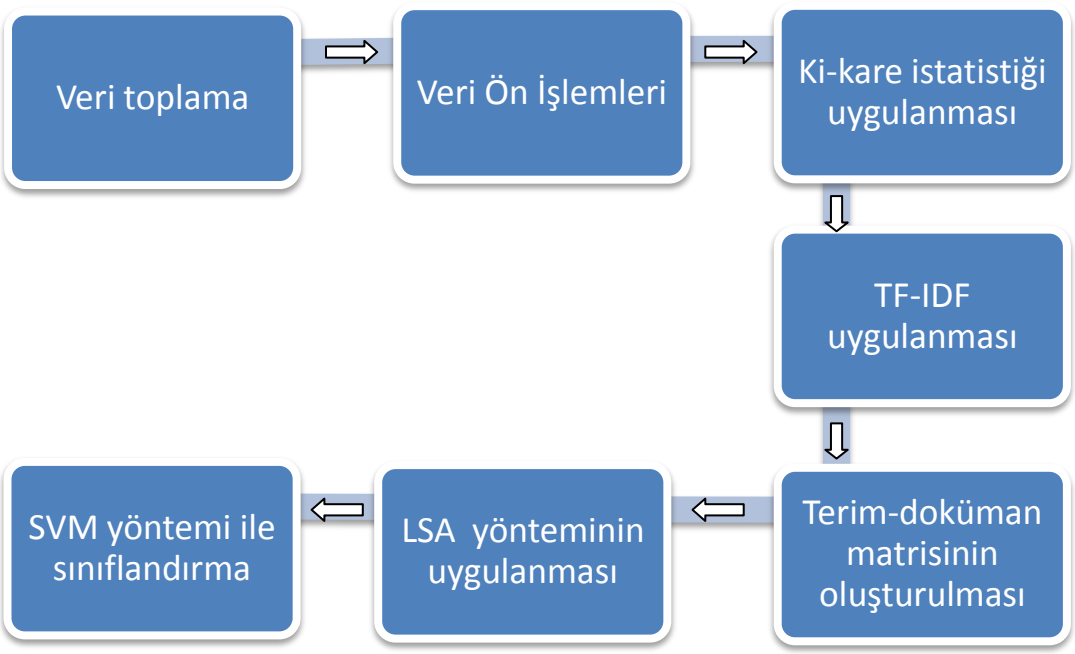
Bu amaç ve hedeflere ulaşmak için web sayfalarının içeriklerine otomatik olarak web sayfası bağlantı adresleri aracılığıyla ulaşılması, içeriğin ayrıştırılması, terimlerin ağırlıklandırılması ve terim indirgenmesi gibi ön işleme gerçekleştirilir. Daha sonra cebirsel ayrıştırma yöntemi olan tekil değer ayrışımının yine internet yazılımıyla gerçekleştirmesi ve sonuçların görsel içerikleriyle sunulması ile yapılan çoğu deneysel çalışmanın bir internet uygulamasına dönüştürülmesi gerçekleştirilmiştir.

3.1. Çalışmanın Önemi

Bilgisayar sistemleri her geçen gün ucuzlamakta, saklama kapasiteleri artmakta ve ulaşımı kolaylaşmaktadır. Bunun sonucu olarak büyük boyutlardaki veriler sayısal ortamlarda tutulmaktadır. Ayrıca, bilgiye ulaşmanın hızlı, ucuz ve kolay olmasının yanında bilgisayar ağ sistemleri ve uygulamaları hızla yayılmaktadır. Sonuçta, internet üzerindeki veriler gün geçtikçe büyük bir ivmeyle katlanarak büyümekte, bilgisayar sistemleri üzerinde tutulan bu devasa veri yığınları her geçen gün organize ve optimizasyon edilmeye gerek duymaktadır. Bu yapılandırılmamış veri yığınlarının yapılandırılmış veriye dönüştürülmesi, özetlenmesi, yararlı bilgilerin çıkarılması gibi çalışmalara ihtiyaç duyulmaktadır.

Yukarıda bahsettiğimiz hedefler doğrultusunda bu çalışma soncu önerdiğimiz sınıflandırma algoritması veri yığınlarının sınıflandırılmasına yardımcı olacaktır.

Algoritmanın temel aşamalarını gösteren akış şeması Şekil 3.1' de verilmiştir



Şekil 3.1. Geliştirilen Algoritmanın Temel Aşamaları

Şimdi bu temel aşamalarda ne tür işlemlerin yapıldığını detaylandıralım.

3.2. Veri Toplama Süreci

Birçok web içerik madenciliği çalışmasında internet sayfalarının içerikleri sayfa kaynaklarının çıkarılması yöntemiyle toplanır. Bu çalışmada ise, uygulamada kullanılmak istenen web sayfalarının içeriklerine sadece adres bilgileriyle ulaşılmaktadır.

Uygulamanın yazılmış olduğu PHP programlama dilinin cURL kütüphanesi yardımıyla farklı web sunuculardaki içeriklere ulaşılır (Bkz. EK 1).

3.2.1. CURL Kütüphanesi

PHP programlama dili bize farklı tip sunucu ve farklı tip protokole bağlanmamıza ve iletişime geçmemize olanak sağlayan cURL kütüphanesini desteklemektedir. cURL son olarak http, https, ftp, gopher, telnet, dict, file ve Idap protokollerini desteklemektedir. cURL ayrıca "http post", "http put", "ftp uploading" ve "http form temelli upload", vekil sunucular, çerezler ve kullanıcı-şifre izinlerini de desteklemektedir [32].

3.3. Veri Ön İşleme Süreci

Bilindiği gibi web sayfaları HTML etiketleri, CSS kodları, javascript gibi çeşitli istemci taraflı programlama dilleri kod parçaları ve meta verileri gibi yapılar içerir. Metin madenciliği süreçlerine geçebilmek için öncelikle bu web sayfası içeriği dışındaki bilgilerin temizlenmesi gerekmektedir.

Bu çalışmada veri ön işleme işlemleri olarak aşağıdaki aşamalardan geçilmiştir (Bkz. EK 2).

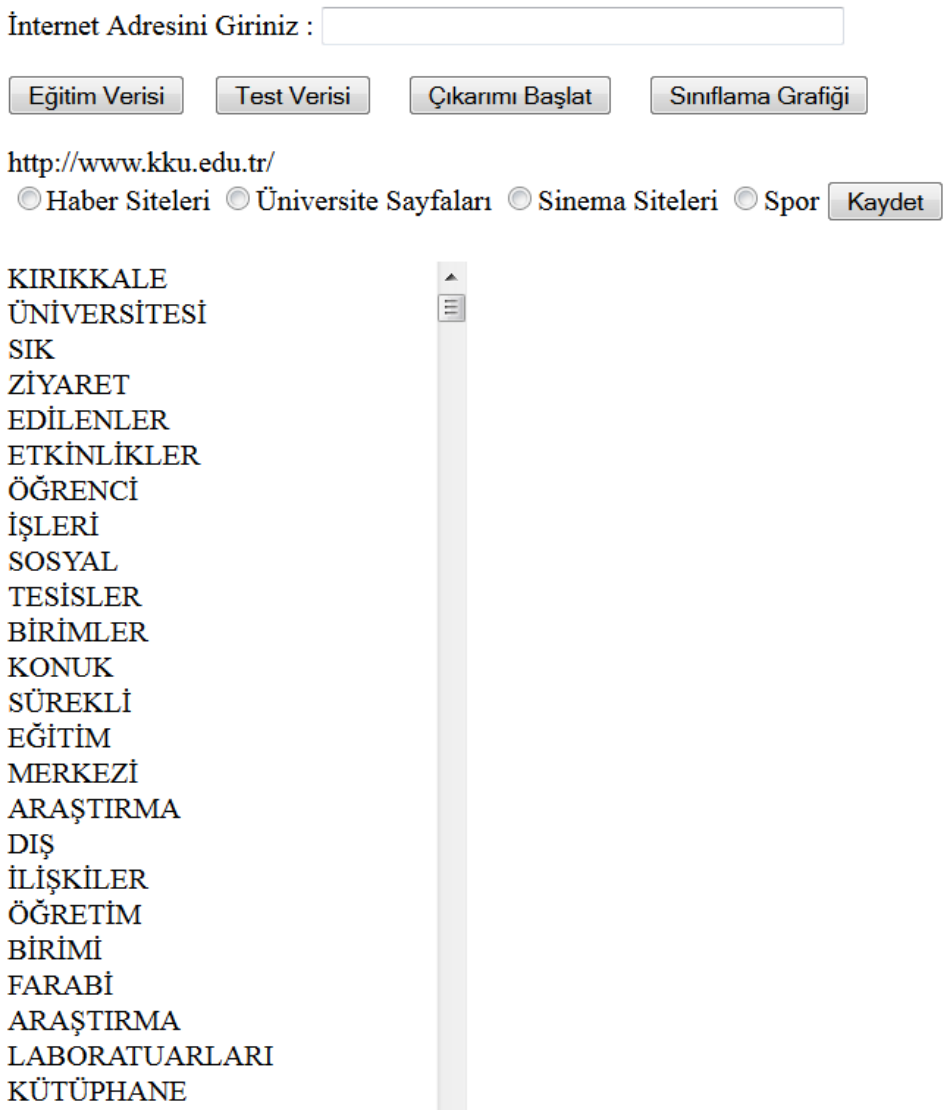
- Tüm içeriğin büyük harfe çevrilmesi
- Html etiketlerinin temizlenmesi
- Boşlukların temizlenmesi
- Sembollerin temizlenmesi
- Noktalama işaretlerinin temizlenmesi
- Rakamların temizlenmesi

Eğitim seti oluşturulduktan sonra web içerikleri incelenerek metinler terimlere ayrıştırılır. Çalışmada değerlendirme yapılacak terim seti eğitim başlamadan önce belli değildir. Günümüz metin sınıflandırma çalışmalarında genellikle terim seti eğitim başlamadan önce belirlenmektedir.

Ayrıca, belli bir sınıf için değerlendirmeye alınacak terimler bir uzman tarafından seçilebilir veya en sık kullanılan kelime kümesi kullanılabilir. Ancak Web içeriklerinin sınıflandırılmasında bu metotlar her zaman iyi sonuç vermeyebilir. Web içerikleri dinamiktir ve belli zaman aralıklarla değişip güncellenebilir. Örneğin spor sayfalarının sınıflandırıldığı bir çalışmada Avrupa Basketbol Turnuvası'nın oynandığı bir dönemde yapılan çalışmanın terim havuzunda basketbol, fiba, oniki dev adam gibi terimler veya turnuvada yıldızı parlayan oyuncular ve takımların isimlerinin veri setinde olmasını isteriz, çünkü spor sayfalarında bu terimler sık geçeceği için ayırt etmede büyük fark yaratacaktır. Bu örnekten de anlaşılabileceği gibi bilginin değişip geliştiği bir ortamda sabit veri seti kullanmak sınıflandırma başarısını düşürebilir. Bu çalışmada terimler sınıflandırmaya dahil olan metin içeriklerinden çıkarılmaktadır.

Terimlerin seçiminde sınıflama algoritmasının başarısını yüksek tutmak için en sıklıkla geçen ve sınıflar için en ayırt edici olan terimleri seçme gayesiyle çeşitli yöntemler uygulanmıştır. Bunlar, frekansı çok düşük olan kelimelerin çıkarılması ve ki-kare test skorları yeterli eşik değerini geçmeyen terimlerin çıkarılmasıdır.

Şekil 3.2'de Kırıkkale Üniversitesi internet sayfasının eğitim setine alınırken içeriğin elde edilmesi ve kelime ayrıştırmasından bir kısım gösterilmiştir. İnternet sayfasından elde edilen kelimeler resmin sol alt kısmında yer almaktadır. Elde edilen sayfa kelimeleri veritabanında bir havuzda toplanacaktır.



Şekil 3.2 Eğitim Setine Katılmak Üzere İçeriğe Erişme ve Ayrıştırma İşlemi

3.3.1. Ki Kare Testinin Uygulanması

Terim-doküman matrislerinin büyük çoğunluğu seyrek matrislerdir. Bu durum işlem süresi ve saklama maliyeti ihtiyacını yükseltmektedir. Ki-kare testi, en faydalı terimleri seçer ve nispeten daha az faydalı terimleri temizler. Bu işlemi yaparken terimlerin ki-kare skoru değerinden faydalanır. Çalışmada çeşitli ki-kare skorları seçilip başarının değişimi izlenmiştir.

Çizelge 3.1'de Ki-kare testi uygulamasından sonra uygulamanın hesapladığı ki-kare skorlarına göre sıralanmış terimlerin bir kısmı tablo halinde gösterilmiştir.

Çizelge 3.1. Ki-Kare Test Sonuçları

Eğitim Kelimeler İstatistikleri			
Terim	**Terim Ki-Kare Skoru**	**Varlığından Dolayı Ayırıcılığı**	**Yokluğundan Dolayı Ayırıcılığı**
İDARİ	22.577	Üniversite Sayfaları	Haber Siteleri
AKADEMİK	20.9473	Üniversite Sayfaları	Haber Siteleri
ERDOĞAN	19.0336	Haber Siteleri	Üniversite Sayfaları
KÜTÜPHANE	18.9605	Üniversite Sayfaları	Haber Siteleri
PARKI	18.9296	Haber Siteleri	Üniversite Sayfaları
ÜNİVERSİTEMİZ	17.4267	Üniversite Sayfaları	Haber Siteleri
GEZİ	16.9624	Haber Siteleri	Spor
BAŞBAKAN	16.2767	Haber Siteleri	Üniversite Sayfaları
PERSONEL	16.0921	Üniversite Sayfaları	Haber Siteleri
BİRİMLER	15.6183	Üniversite Sayfaları	Haber Siteleri
POLİS	15.3161	Haber Siteleri	Üniversite Sayfaları
VİZYONDA	15.2215	Sinema Siteleri	Haber Siteleri
FİKSTÜR	14.8415	Spor	Haber Siteleri
FAKÜLTESİ	14.4621	Üniversite Sayfaları	Haber Siteleri
VOLEYBOL	14.4373	Spor	Haber Siteleri
ETKİNLİKLER	14.4201	Üniversite Sayfaları	Haber Siteleri
BAKAN	14.2906	Haber Siteleri	Sinema Siteleri
HABERLERİ	14.1955	Haber Siteleri	Üniversite Sayfaları
ARAŞTIRMA	14.1776	Üniversite Sayfaları	Haber Siteleri
ÖĞRETİM	14.0589	Üniversite Sayfaları	Haber Siteleri
DAİRE	14.0338	Üniversite Sayfaları	Haber Siteleri
AÇIKLAMA	13.8318	Haber Siteleri	Üniversite Sayfaları
TAKSİM	13.5866	Haber Siteleri	Üniversite Sayfaları
BASKETBOL	13.4649	Spor	Haber Siteleri
TRABZONSPOR	13.429	Spor	Üniversite Sayfaları

Çizelge 3.1‘de web içerikleri üzerinden çıkarılan kelimeler ki-kare skoruna göre bir kısmının azalan bir şekilde sıralanmıştır. Bu tabloda ki-kare skorlarının yanında birde hangi sınıf için belirleyici olduğu gösterilmiştir. Örneğin ki-kare skoruna göre birinci sırada yer alan “idari” kelimesi üniversite sayfalarında sıklıkla geçtiği ve diğer sınıflarda bulunmadığı veya az bulunduğu için bu sınıf için varlığından dolayı ayırıcıdır. “idari” kelimesi haber sitelerinde ise hiç bulunmaması veya çok az bulunmasından dolayı bu terimin haber siteleri için yokluğundan dolayı ayırıcıdır denir ve bu iki durumun aynı anda diğerlerine göre en üst düzeyde olması en yüksek ki-kare skoruna sahip olmasını sağlamıştır.

Çizelge 3.1‘de koyu olarak gösterilen kelimeler, Şekil 3.2’de bir üniversite sayfasının terimlerinin ayrıştırılması sırasında gösterilen kelime grubunda ortak bulunanların bazılarıdır ve dikkat çekmek amacıyla koyu fonttadır. Bunlar; kütüphane, birimler, etkinlikler ve öğretim kelimeleridir. Bu kelimelerin ki-kare skorları yüksek olduğundan değerlendirilme yapılacak terim kümesine alınırlar ve sınıflandırma başarısını artırırlar.

3.4. Terim Ağırlıklandırma

Özellik ağırlandırma yöntemi bir terim doküman vektörünü oluştururken her bir terim için kullanılan değerin sadece varlık veya yokluk dışında daha hassas değerler vererek sınıflandırmaya yardımcı olabilmektedir. Bu çalışmada ayrıca terim frekansı–ters doküman frekansı ağırlıklandırma yöntemi uygulanmıştır, böylece daha hassas bir derecelendirme yaparak sınıflandırma başarım oranları artırılmıştır.

3.5. Terim-Doküman Matrisinin Oluşturulması

Veri ön işlem sürecinden sonra değerlendirmede kullanılacak terimler belirlenmiştir. Kelimeleri indirgemede kullanılan dokümanlar ve test için kullanılacak dokümanların her biri için terim vektörü oluşturulur. Bundan sonra

her bir doküman için terim-doküman vektörleri oluşturulur ve terim-doküman matrisleri elde edilerek matematiksel ortama taşınır.

3.6. SVD Uygulanması

Terim-doküman matrisinin oluşturulmasından sonra LSA'nın ilk aşaması olan terim-doküman matrisinin tekil değer ayrışımı hesaplanır ikinci aşamada ise tekil değerlerin büyüklüğüne bakılarak matris boyutu indirgenir. Bu indirgeme işleminde belli bir eşik değerinden düşük tekil değerler ve buna karşılık gelen tekil vektörler göz ardı edilir. Sonuçta orijinal doküman terim matrisini iyi temsil eden daha ekonomik boyutlara sahip bir doküman terim matrisi elde edilir.

Doküman vektörleri yardımıyla dokümanlar bir koordinat sistemine yerleştirilebilir. Şekil 3.3'de dokümanların koordinat sistemi üzerinde konumlandırıldığı uygulama ekranı gösterilmiştir.

Şekil 3.3. Dokümanların Gizli Anlam Analizi Sonucunda Sınıflandırılması.

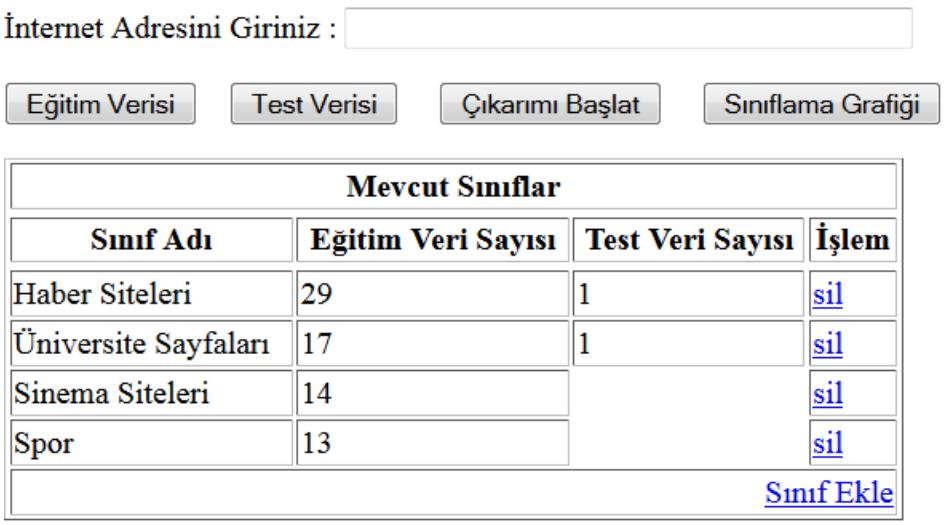
3.7. Destek Vektör Makinesi Yöntemiyle Başarı Değerlendirmesi

Makine öğrenmesinde sınıflandırma problemlerinde kullanılan bir yöntem olan Destek Vektör Makinesi veriyi sınıflandırarak birbirinden ayırmak için en uygun fonksiyonun tahmin edilmesi esasına dayanır. Bu çalışmada LSA'dan elde edilen doküman vektörler destek vektör makinesi yöntemiyle sınıflandırılmıştır. Metin sınıflama çalışmalarında her ne kadar destek vektör makinesinin tahminin başarıya etkisi olsa da asıl önemli işlem sınıflandırma yapılacak dokümanın sınıflarını birbirinden daha kolay ayrıştırılmasını sağlayacak terim-doküman matrisinin elde edilmesidir.

3.8. Uygulama Genel Adımları

Bu çalışmanın yayınlandığı ip adresi http://193.255.242.10/index.php'dir. Bu adresten uygulama giriş sayfasına ulaşılabilir.

Şekil 3.4'de uygulama giriş ekranı gösterilmiştir.



Şekil 3.4. Uygulama Girişi Ekran Görüntüsü

Uygulama kullanım adımları aşağıdaki gibidir.

1. Eğitim setinde kullanılmak istenen internet sayfasının bağlantı (link) adresi internet adresi giriş alanına yazılır.
2. Eğitim verisi butonu yardımıyla eğitim seti oluşturulmaya başlanır. Her bir internet adresi kaydında birinci ve ikinci adım tekrarlanır. Alt kısımdaki mevcut sınıflar tablosunda yapılan kayıtların sayıları listelenir.
3. Eğitim kümesi oluşturulduktan sonra artık test için kullanılacak olan web sayfası adresleri yine internet adres kutusuna girilir. Ancak test seti oluşturulacağından "Test Verisi" butonu tıklanarak kayıt işlemi yapılır.
4. Eğitim ve test setleri oluşturulduktan sonra "Çıkarımı Başlat" butonu yardımıyla terim seti, doküman vektörleri ve terim-doküman matrisi hazır hale getirilir.
5. Son adımda ise artık eğitim ve test başarılarının görsel ve sayısal sonuçları "Sınıflama Grafiği" butonu yardımıyla görülür.

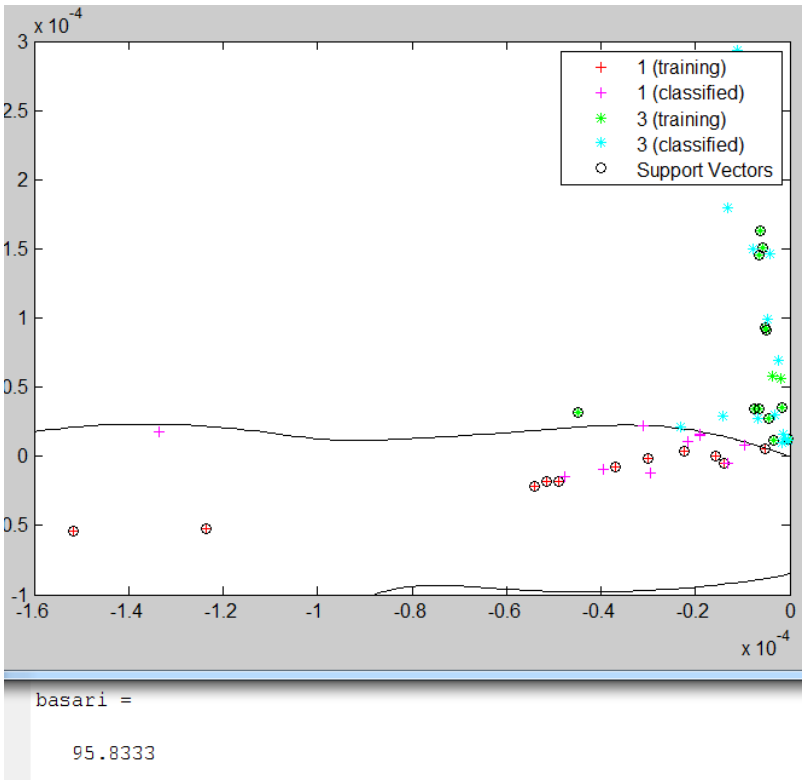
4. ARAŞTIRMA BULGULARI

Bu çalışmada, gerçekleştirilen uygulama yazılımıyla web içeriklerinin gizli anlam analizi yöntemi ile değerlendirilmesi ve destek vektör makineleri sınıflandırma yöntemi kullanılmıştır. Gerçeklenen yazılım ile 2, 3 ve 4 kategoride eğitim kümeleri oluşturulmuş 100 doküman ile eğitimler yapılmıştır. Test başarımını artırmak için uygulanan yöntemler analiz edilmiş ve başarıya etkileri çıkarılmıştır.

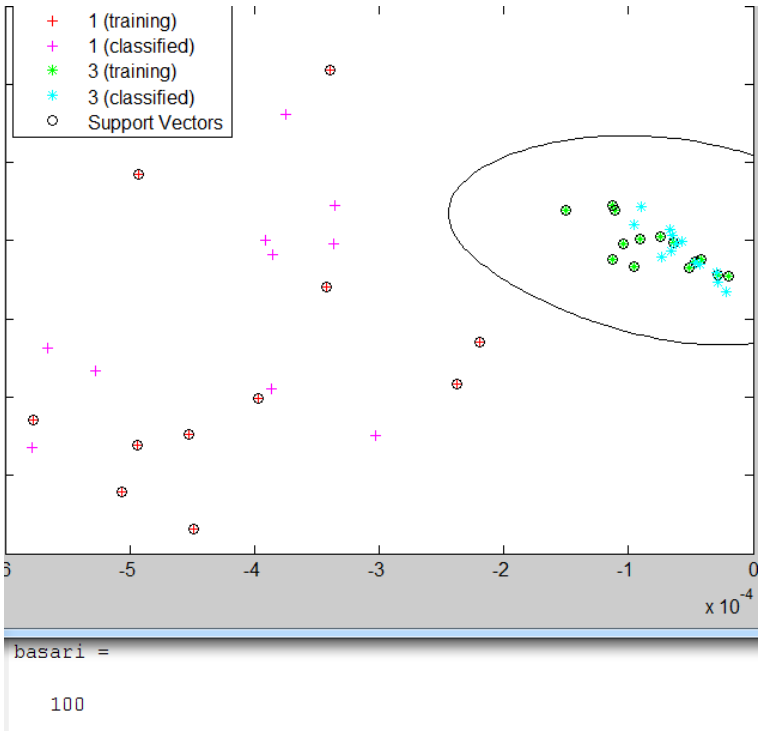
Yapılan çalışmada haber sayfası, spor sayfası, sinema sayfası ve üniversite sayfası olmak üzere toplam 4 sınıf kullanılmıştır. Elde edilen sonuçlardan ve uygulama çalışması boyunca yapılan gözlemlerden sınıf sayısı arttıkça başarı oranının nispeten azaldığı sonucuna varılmıştır.

Çalışmada terim ağırlıklandırma ve terim indirgeme yöntemlerinin başarıya etkileri de test edilmiştir. Yapılan çalışmalarda sınıflar birbiriyle ilişkili sınıflar seçildiğinde en düşük %85 başarı iken birbiriyle daha az ilişkili sınıflar seçildiğinde başarı oranı %100 artmıştır. Örneğin haber ve spor sayfa sınıfları birbiriyle ilişkilidir, yani birçok haber sayfası içerik olarak çok büyük bir kısım olmasa da spor haberlerini içerir. Spor ve üniversite sayfaları düşünüldüğünde iki sınıf birbiriyle oldukça az ilişkilidir ve yapılan çalışmalarda bu sınıflar arasındaki sınıflandırma başarısı %100 olarak bulunmuştur.

Şekil 4.1'de haber ve spor sayfalarının destek vektör makinesi yöntemi kullanılarak yapılan sınıflandırma sonuçları gösterilmiştir. Yeşil ile gösterilen dokümanlar haber sitesi eğitim dokümanları, kırmızı renk ile gösterilen dokümanlar ise spor sayfası eğitim dokümanlardır. Mavi renkteki haber verileri ve pembe renkteki spor verileri ise test amaçlı kullanılmıştır. Başarı oranı %95,83 bulunmuştur. Destek vektör makinesi sınıflandırma yöntemi ile çizilen uygunluk fonksiyonu eğrisi iki sınıfı başarıyla ayırmıştır. Gözlemleyelim ki yanlış değerlendirilen dokümanlar eğri sınırının üzerinde bulunmaktadır.



Şekil 4.1. Haber ve Spor Sayfalarının Sınıflandırma Sonuçları



Şekil 4.2. Haber ve Spor Sayfası tf-idf Uygulanmış Sınıflandırması

Şekil 4.2'de Yapılan ikinci deneyde haber ve spor sınıfı terimlerine terim frekansı–doküman ters frekansı ağırlıklandırma yöntemi uygulanmış ve tüm test verilerini doğru olarak sınıflandırmıştır. Başarı oranı %95,83'den %100'e getirerek terim ağırlıklandırmanın başarımı tam olarak yükselttiği gözlemlenmiştir.

Sınıflandırma başarısını etkileyen bir diğer faktör ise doğru terimlerin seçilmesi yanında etkisiz terimlerin temizlenmesidir.

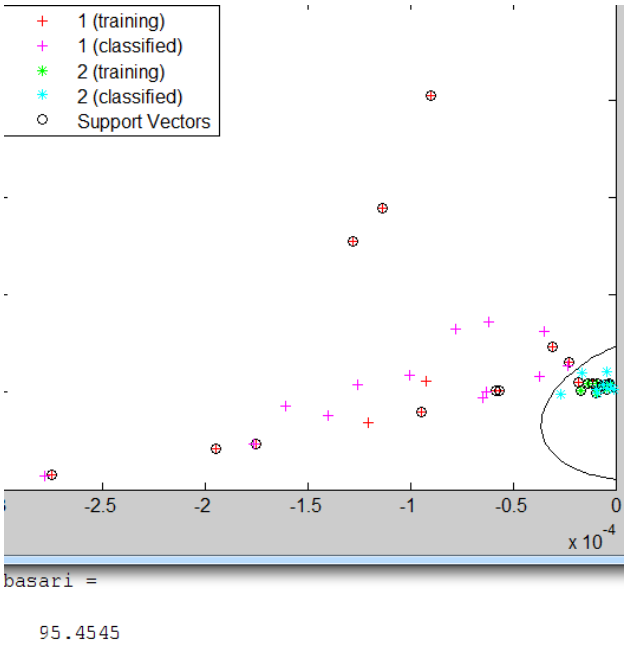
Bu çalışmada ilk olarak dokümanlarda geçme oranı seyrek olan terimlerin temizlenme adımı yapılmıştır. Genellikle bu kelimeler teknik veya kullanımı az olan kelimelerdir. Örneğin "müphem" gibi kullanımı az olan kelimeler ya da "plonjon" ve "futsal" gibi spor terimleri kendi spor sınıflarında bile nadiren bulunabileceğinden bu terimlerin uygulama tarafından sıklığı az bulunarak otomatik olarak temizlenmesi yapılır.

İkinci olarak terimlerin sınıfları ayırt ediciliğinin bir araştırması olan ki-kare testi uygulanmıştır. Yeterli ayırt ediciliği olmadığı tespit edilen kelimeler terim havuzundan çıkarılmıştır. Yani, bazı terimler her bir sınıfta dengeli şekilde bulunuyor olabilir bu durumda bu terimlerin sınıflamada ayırt edici bir özelliğe sahip olmayacaktır. Örneğin, "birlikte " ve "ayrıca" gibi kelimeler her sınıf içinde bulunabileceğinden değerlendirmeye alınarak işlem gücü ve saklama maliyetlerini artıracak ve başarımı düşürecektir. Geliştirilen uygulama bu durumu da dikkate alarak bu türdeki terimleri ayıklama özelliğine sahiptir. Yapılan çalışma sonucunda kelime havuzundan çıkarılan bazı kelimeler Çizelge 4.1'de verilmiştir.

Çizelge 4.1. Ki-kare Test Skoru Düşük Bulunan Bazı Kelimeler

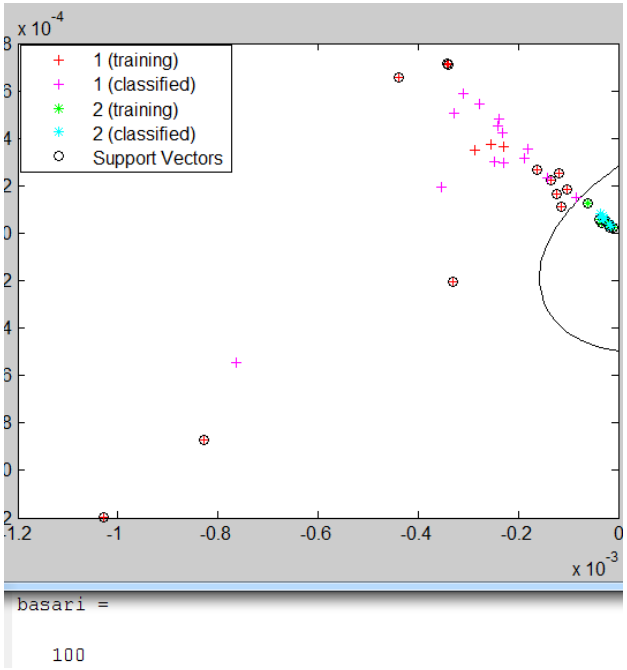
türk	genel
önemli	türkçe
fotoğraf	tarafından
saat	büyük
gece	destek

Şekil 4.3'de öncelikle haber ve üniversite sayfaları arasındaki sınıflandırma çalışması sonuçları gösterilmiştir. Bu çalışmada 50 adet internet sayfa kümesi kullanılmış toplam 350 kelime ile çalışılmış ve %95,45 oranında başarı sağlanmıştır.



Şekil 4.3. Haber ve Üniversite Sayfaları Sınıflama Sonucu

Şekil 4.4 ‘de ise aynı internet sayfası kümesinden çıkarılan 350 kelimeye ki-kare testi uygulandı ve 130 kelime test sonucunda yeterli ayırt ediciliği sağlayamadığı için temizlenmiş geriye kalan 230 kelime ile çalışma yapılmıştır. Uygulama sonucunda %100 başarı sağlanmıştır.



Şekil 4.4. Haber ve Üniversite Sayfaları Ki-kare Yöntemiyle Kullanımı

5. TARTIŞMA VE SONUÇ

Teknolojinin hızlı gelişimi ve ucuzlaması ile birlikte günümüzde bilgiye erişim eskiye oranla daha çok sayısal belgeler üzerinden olmaktadır. Özellikle internetin hızlı bir şekilde yaygınlaşması, hayatımızı kolaylaştırmış ve bilgiye erişim, transfer ve saklanma yöntemlerimizi değiştirmiştir. Artık insanlar internet sayesinde sadece bilgiye erişmemekte, bilgi oluşturup yine internet ortamında paylaşabilmektedir. Bu özgür ortamda veri miktarları her geçen gün hızla artmakta, devasa veri yığınları oluşturmaktadır.

Veri boyutlarındaki bu hızlı artış veriye erişim ve veriyi bilgiye dönüştürme problemini de beraberinde getirmiştir. Veriler düzensiz bir yapıda depolanmakta bu da verilerin değerini düşürmektedir. Çünkü yapılandırılmamış verilere istenildiğinde ulaşmak, bu bilgilerden ihtiyaç duyulan bilgilerin çıkarılmak oldukça güç olmaktadır. Bu noktada veri madenciliği çalışmaları ortaya çıkmış ve değerli bilgilerin hızlı bir şekilde gün yüzüne çıkarılması, verilerin özetlenmesi, sınıflandırılması ve daha birçok konuda çözümler üretilmiştir.

Veri madenciliği üzerinde yapılan çalışmalar genel olarak veritabanları gibi yapısal veriler üzerine odaklanmıştır. Ancak günümüzde erişilebilir ve kullanılabilir durumdaki verinin önemli bir kısmı doküman ve dosyalar üzerinde bulunmaktadır. Bunlar genel olarak makaleler, elektronik postalar, kitaplar, sayısal kütüphaneler ve web sayfaları gibi çok geniş doküman türleri olabilir. Metin madenciliği ise yararlı ve kaliteli verinin bilgi işlem yöntemleriyle yapısal olmayan yani metin halindeki veriden elde edilmesidir. Sınıflandırma, bilgi çıkarımı, kümeleme ve birliktelik analizi gibi birçok metin madenciliği tekniği bulunmaktadır. Metinleri, önceden tanımlanmış sınıflara atama işlemine metin sınıflandırma denilmektedir. Metin sınıflandırma, arama motorları gibi veriye daha kolay erişme yöntemleri için hayati önem taşımaktadır ve literatürde metin sınıflandırılması konusunda önemli çalışmalar bulunmaktadır. Ancak teknolojinin hızlı gelişmesi ve kullanımının yaygınlaşması bu alanda yapılan çalışmaların yetersiz kalmasına neden olmaktadır.

İnternetin gelişmesiyle metinlerle birlikte daha birçok yapıdaki veri sayısal ortama taşınmıştır. İnternet sayesinde mekandan bağımsız ve çok ucuza veriye hızlı ulaşılabilmekte, daha fazla veri saklanabilmekte ve paylaşılabilmektedir. İnternet üzerindeki verinin hiç olmadığı kadar hızla büyümesi ile web üzerindeki verilerde bilgi keşfi sürecinde veri madenciliğinin alt dalı olan web madenciliği ortaya çıkmıştır. Metin madenciliği konusu olan metin sınıflandırma çalışmaları benzer şekilde internet üzerinde yapılabilmekte ve web madenciliğinin alt kategorisi olan web içerik madenciliğini çalışmaları olarak bilinmektedir. İnternetin barındırdığı gürültülü bilginin çeşitliliği ve farklı veri yapıları nedeniyle web sayfalarının içerik sınıflandırması doğal metin sınıflandırmasına göre daha karmaşık ve zordur.

Web içeriklerinden anlamlı bilgilerin çıkarılması, bilginin özetlenmesi, bilginin sınıflandırılması yani yapısal olmayan verinin yapısal veriye dönüştürülmesi önemli çalışma alanları olmuştur. Özellikle internet sayfalarında bulunan içeriklerin dinamik yapıda olması, sayfa içeriklerinin gün içerisinde bile onlarca kez güncellendiği düşünüldüğünde başarılı bir sınıflandırma yapabilmek önemli bir çalışma konusudur.

Çalışmamızda web sayfalarının otomatik sınıflandırılması için bir algoritma geliştirilmiştir. Yapılan çalışma sonucunda veri ön hazırlama sürecinin sınıflandırma başarısında çok önemli olduğu tespit edilmiştir. Terim indirgeme işlemi yapılmadan önce belli bir test grubunda $\%85,84$ başarı oranı sağlanırken terim indirgeme yapıldıktan sonra başarı oranı $\%100$'e kadar çıktığı görülmüştür. Bir diğer önemli veri ön hazırlama sürecide terim ağırlıklandırma yöntemi olmuştur. Terim ağırlığı olarak terim frekansı yöntemiyle yapılan bir çalışmada $\%95.83$ olan sınıflandırma başarısı, tf-idf ağırlıklandırma yöntemiyle $\%100$ olarak bulunmuştur.

Yapılan sınıflandırma çalışmalarında $\%85 - \%100$ arasında sınıflandırma başarısı sağlanmıştır.

İleriki çalışmalarda tezde geliştirilen sınıflandırma algoritmasının başarısını artırmak için daha iyi bir terim ağırlıklandırma yöntemi ve özellik seçim yöntemi üzerinde çalışma yapmayı hedeflemektedir.

EK 1. CURL KÜTÜPHANESİNİN KULLANIM KODU

```
$url = "http://www.ntvmsnbc.com/";
$ch = curl_init(); // cURL oturumu başlatılıyor.
curl_setopt($ch,CURLOPT_HTTPHEADER,
    array (
        "Content-Type: application/x-www-form-urlencoded; charset=utf-8"
    )
);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, 1); // Sayfa içeriğini döndür.
curl_setopt($ch, CURLOPT_URL, $url);  // girilen url bilgisini ata
$result = curl_exec($ch);
$info = curl_getinfo($ch);
curl_close($ch); // curl kaynaklarını kapat ve sistem kaynaklarını serbest bırak
```

EK 2. VERİ ÖN İŞLEME AŞAMALARI

Büyük harfe çevirme fonksiyonu

```
function strtouppertr($metin){
   return mb_convert_case(str_replace('i','İ',$metin), MB_CASE_UPPER, "UTF-8");
}
```

HTML etiketlerini temizleme fonksiyonu

```
function strip_html_tags($str) {
   $str = preg_replace('/(<|>)\1{2}/is', '', $str);
   $str = preg_replace(
      array(
         '@<style[^>]*?>.*?</style>@siu',
         '@<script[^>]*?.*?</script>@siu',
         '@<noscript[^>]*?.*?</noscript>@siu',
         ),
      " ",
      $str );
   $str = replaceWhitespace($str);
   $str = strip_tags($str);
   return $str;
}
```

Boşlukları temizleme fonksiyonu

```
function replaceWhitespace($str) {
   $result = $str;
   foreach (array(
   "  ", " \t",  " \r",  " \n",
   "\t\t", "\t ", "\t\r", "\t\n",
   "\r\r", "\r ", "\r\t", "\r\n",
```

```
    "\n\n", "\n ", "\n\t", "\n\r",
    ) as $replacement) {
    $result = str_replace($replacement, $replacement[0], $result);
    }
    return $str !== $result ? replaceWhitespace($result) : $result;
}
```

Sembolleri temizleme fonksiyonu

```
function strip_symbols( $text )
{
        $plus   = '\+\x{FE62}\x{FF0B}\x{208A}\x{207A}';
        $minus  = '\x{2012}\x{208B}\x{207B}';

        $units  = '\\x{00B0}\x{2103}\x{2109}\\x{23CD}';
        $units .= '\\x{32CC}-\\x{32CE}';
        $units .= '\\x{3300}-\\x{3357}';
        $units .= '\\x{3371}-\\x{33DF}';
        $units .= '\\x{33FF}';

        $ideo   = '\\x{2E80}-\\x{2EF3}';
        $ideo  .= '\\x{2F00}-\\x{2FD5}';
        $ideo  .= '\\x{2FF0}-\\x{2FFB}';
        $ideo  .= '\\x{3037}-\\x{303F}';
        $ideo  .= '\\x{3190}-\\x{319F}';
        $ideo  .= '\\x{31C0}-\\x{31CF}';
        $ideo  .= '\\x{32C0}-\\x{32CB}';
        $ideo  .= '\\x{3358}-\\x{3370}';
        $ideo  .= '\\x{33E0}-\\x{33FE}';
        $ideo  .= '\\x{A490}-\\x{A4C6}';

        return preg_replace(
                array(
```

```
'/[\p{Sk}\p{Co}]/u',
'/\p{Sm}(?<![' . $plus . $minus . '=~\x{2044}])/u',
'/((?<= )|^)[' . $plus . $minus . ']+((?![\p{N}\p{Sc}])|$)/u',
'/((?<= )|^)=+/u',
'/[' . $plus . $minus . '=~]+((?= )|$)/u',
'/\p{So}(?<![' . $units . $ideo . '])/u',
'/ +/',
),
' ',
$text );
}
```

Noktalama işaretleri temizleme fonksiyonu

```
function strip_punctuation( $text )
{
   $urlbrackets    = '\[\]\(\)';
   $urlspacebefore = ':;\'_\*%@&?!' . $urlbrackets;
   $urlspaceafter  = '\.,:;\'\-_\*@&\/\\\\\?!#' . $urlbrackets;
   $urlall         = '\.,:;\'\-_\*%@&\/\\\\\?!#' . $urlbrackets;
   $specialquotes  = '\'"\*<>';
   $fullstop       = '\x{002E}\x{FE52}\x{FF0E}';
   $comma          = '\x{002C}\x{FE50}\x{FF0C}';
   $arabsep        = '\x{066B}\x{066C}';
   $numseparators  = $fullstop . $comma . $arabsep;
   $numbersign     = '\x{0023}\x{FE5F}\x{FF03}';
   $percent        =
'\x{066A}\x{0025}\x{066A}\x{FE6A}\x{FF05}\x{2030}\x{2031}';
   $prime          = '\x{2032}\x{2033}\x{2034}\x{2057}';
   $nummodifiers   = $numbersign . $percent . $prime;
   return preg_replace(
      array(

         '/[\p{Z}\p{Cc}\p{Cf}\p{Cs}\p{Pi}\p{Pf}]/u',
```

```
            '/\p{Po}(?<![' . $specialquotes .
               $numseparators . $urlall . $nummodifiers . '])/u',
            '/[\p{Ps}\p{Pe}](?<![' . $urlbrackets . '])/u',
            '/[' . $specialquotes . $numseparators . $urlspaceafter .
               '\p{Pd}\p{Pc}]+((?= )|$)/u',
            '/((?<= )|^)[' . $specialquotes . $urlspacebefore . '\p{Pc}]+/u',
            '/((?<= )|^)\p{Pd}+(?![\p{N}\p{Sc}])/u',
            '/ +/',
        ),
        ' ',
        $text );
}
```

Rakamları temizleme fonksiyonu

```
function strip_numbers( $text )
{
    $urlchars      = '\.,:;\'=+\-_\*%@&\/\\\\?!#~\[\]\(\)';
    $notdelim      = '\p{L}\p{M}\p{N}\p{Pc}\p{Pd}' . $urlchars;
    $predelim      = '((?<=[^' . $notdelim . '])|^)';
    $postdelim     = '((?=[^'  . $notdelim . '])|$)';

    $fullstop      = '\x{002E}\x{FE52}\x{FF0E}';
    $comma         = '\x{002C}\x{FE50}\x{FF0C}';
    $arabsep       = '\x{066B}\x{066C}';
    $numseparators = $fullstop . $comma . $arabsep;
    $plus          = '\+\x{FE62}\x{FF0B}\x{208A}\x{207A}';
    $minus         = '\x{2212}\x{208B}\x{207B}\p{Pd}';
    $slash         = '[\/\x{2044}]';
    $colon         = ':\x{FE55}\x{FF1A}\x{2236}';
    $units         = '%\x{FF05}\x{FE64}\x{2030}\x{2031}';
    $units        .= '\x{00B0}\x{2103}\x{2109}\x{23CD}';
    $units        .= '\x{32CC}-\x{32CE}';
    $units        .= '\x{3300}-\x{3357}';
```

```
    $units        .= '\x{3371}-\x{33DF}';
    $units        .= '\x{33FF}';
    $percents      = '%\x{FE64}\x{FF05}\x{2030}\x{2031}';
    $ampm          = '([aApP][mM])';
    $digits        = '[\p{N}' . $numseparators . ']+';
    $sign          = '[' . $plus . $minus . ']?';
    $exponent      = '([eE]' . $sign . $digits . ')?';
    $prenum        = $sign . '[\p{Sc}#]?' . $sign;
    $postnum       = '([\p{Sc}' . $units . $percents . ']|' . $ampm . ')?';
    $number        = $prenum . $digits . $exponent . $postnum;
    $fraction      = $number . '(' . $slash . $number . ')?';
    $numpair       = $fraction . '([' . $minus . $colon . $fullstop . ']' .
       $fraction . ')*';
    return preg_replace(
       array(
          '/' . $predelim . $numpair . $postdelim . '/u',
          '/ +/u',
       ),
       ' ',
       $text );
}
```

KAYNAKÇA

[1] Routledge, F. W., Theories of the Information Age, 2006, s. 30.

[2] Castells, M., The Information Age: Economy, Society And Culture Volume 1-3, 1999.

[3] Gates, B., "Shaping the Internet Age". http://www.microsoft.com/en-us/news/exec/billg/writing/shapingtheinternet.aspx. (Erişim Tarihi: 20.07.2013).

[4] Zhang, W., A comparative study of TF*IDF, LSI and multi-words for text classification, Expert Systems with Applications, s 2758-2765, 2011.

[5] Wei, P., Yang, C., Lin, C., A Latent Semantic Indexing-based approach to multilingual document clustering, Decision Support Systems, s. 606-620, 2008.

[6] Shima, K., SVM-based feature selection of latent semantic features, Pattern Recognition Letters, no. 25, s. 1051-1057, 2004.

[7] Meng, J., Lin, H. A two-stage feature selection method for text categorization, Computers & Mathematics with Applications, s. 2793-2800, 2011.

[8] Yan, H., Augmenting the power of LSI in text retrieval: Singular value rescaling, Data & Knowledge Engineering, s. 108-125, 2008.

[9] Bhat, V., Oates, T., Finding aliases on the web using latent semantic analysis, Data & Knowledge Engineering, no. 49, s. 129-143, 2004.

[10] Yu, B., Latent semantic analysis for text categorization using neural network, Knowledge-Based Systems, no. 21, pp. 900-904, 2008.

[11] Özel, Ş.A., A Web page classification system based on a genetic algorithm using tagged-terms as features, Expert Systems with Applications, no. 38, s. 3407–3415, 2011.

[12] Bayer, H., Veri Madenciliğinde Bir Metin Madenciliği Uygulaması, Beykent Üniversitesi Fen Bilimleri Ensitüsü, İstanbul, 2011.

[13] Veri, http://tr.wikipedia.org/wiki/Veri. (Erişim Tarihi: 21.07.2013).

[14] G. Webinar, Technology Trends You Can't Afford to Ignore, Gartner, 01.07.2009.

[15] Big Data, http://en.wikipedia.org/wiki/Big_data (Erişim Tarihi: 25.07.2013).

[16] Weiss, S. M., Indurkhya, N., Text Mining Predictive Methods for Analyzing Unstructured Information ISBN: 978-0-387-95433-2.

[17] Oğuz, B., Metin Madenciliği Teknikleri Kullanılarak Kulak Burun Boğaz Hasta Bilgi Formlarının Analizi, Akdeniz Üniversitesi Sağlık Bilimleri Enstitüsü, Antalya, 2009.

[18] Britannica, Data mining, http://global.britannica.com/EBchecked/topic/1056150/data-mining (Erişim Tarihi: 30.07.2013).

[19] Dhar, V., Data Science and Prediction, http://www.kdnuggets.com/2012/07/data-science-and-prediction-vasant-dhar.html, Mayıs 2012 (Erişim Tarihi: 30.07.2013).

[20] Uçan, Ö., Dijital Kütüphanelerde Veri Madenciliği Uygulamaları: Akdeniz Üniversitesi Merkez Kütüphanesi Örneği, Antalya: Akdeniz Üniversitesi Sosyal Bilimler Enstitüsü, 2010.

[21] Akpınar, H., Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İstanbul Üniversitesi İşletme Fakültesi Dergisi, ULAKBİM, cilt 29(1), 2010.

[22] Etzioni, O., The World Wide Web: quagmire or gold mine, Comm. of ACM, 1996.

[23] XU, G., Zhang, Y., Li, L., Web Mining and Socieal Networking Techniques and Applications, ISBN 978-1-4419-7734-2 Springer, 2011.

[24] Haberal İ,. Veri Madenciliği Algoritmaları Kullanılarak Günlük Erişimlerinin Analizi, Başkent Üniversitesi Fen Bilimleri Enstitüsü , 2007.

[25] Sristava, J., Cooley, R., Deshpande, M., TAN, P., Web Usage Mining: Discovery and Applications of Usage Patterns From Web Data, SIGKDD Explorations, s. 12-23, 2000.

[26] Kesgin, F., Türkçe Metinler için Konu Belirleme Sistemi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 2007.

[27] Güven, A., Türkçe Belgelerin Anlam Tabanlı Yöntemlerle Madenciliği, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, 2007.

[28] Rouka, L., Metinsel Veri Madenciliğinde Bilgisayarlı Çeviriciler, Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü, Trabzon, 2012.

[29] Biricik, G., Metin Sınıflama İçin Yeni Bir Özellik Çıkarımı, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 2011.

[30] Karaca, M. F., Metin Madenciliği Yöntemi İle Haber Sitelerindeki Köşe Yazılarının Sınıflandırması, Karabük Üniversitesi Fen Bilimleri Enstitüsü, Karabük, 2012.

[31] Lue, B., Web Data Mining, Springer, 2011.

[32] PHP Curl Kütühanesi, http://www.php.net/manual/en/intro.curl.php. (Erişim Tarihi: 10.08.2013).