T.C.
İSTANBUL KÜLTÜR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

RESİM TABANLI OSMANLICA BELGELERDE SINIFLANDIRMA

YÜKSEK LİSANS TEZİ

Ramazan PEHLİVAN

1009042002

Anabilim Dalı: Matematik - Bilgisayar

Programı: Matematik - Bilgisayar

Tez Danışmanı: Yrd. Doç. Dr. Levent ÇUHACI

OCAK 2014

T.C.
İSTANBUL KÜLTÜR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

RESİM TABANLI OSMANLICA BELGELERDE SINIFLANDIRMA

YÜKSEK LİSANS TEZİ

Ramazan PEHLİVAN

1009042002

Anabilim Dalı: Matematik - Bilgisayar

Programı: Matematik - Bilgisayar

Tez Danışmanı: Yrd. Doç. Dr. Levent ÇUHACI

OCAK 2014

ÖNSÖZ

Bu çalısma, İstanbul Kültür Üniversitesi Fen Bilimleri Enstitüsü Matematik-Bilgisayar Anabilim Dalı Yüksek Lisans Tezi olarak hazırlanan "Resim Tabanlı Osmanlıca Belgelerde Doküman Sınıflandırma " isimli tezi içermektedir.

Çalışmalarımın her aşamasında bilgi ve deneyimleri ile yardımcı olan ve bana büyük emekleri geçen, kendisinden çok şey öğrendiğim danışmanım Sayın Yrd. Doç. Dr. Levent ÇUHACI'ya ve zaman zaman mesai dışına taşan uzun çalışma saatlerimize sabırla tahammül gösteren pek değerli eşi Sayın Sibel ÇUHACI'ya içtenlikle teşekkür ederim.

Çalışmamın şekillenmesinde emeği olan, karşılaştığım problemlerde özgün fikirlerinden çokça istifade ettiğim Sayın Yrd. Doç. Dr. R. Murat DEMİRER'e, bilgi ve birikimlerini esirgemeden yol almamdaki yardımları için Sayın Doç. Dr. Banu DİRİ ve Sayın Yrd. Doç. Dr. S. Hikmet ÇAĞLAR'a, Başbakanlık Osmanlı Arşivleri Uzmanı ve Osmanlıca hocam Sayın Dr. Mustafa ÇAKICI 'ya teşekkür ederim.

Yoğun iş tempomun haricinde onlara ait olduğuna inandığım vakitlerden çalarak tez çalışmamı tamamladığım için bana anlayış ve sabır gösteren, onun da ötesinde destek olan sevgili eşim Emel PEHLİVAN, çok sevdiğim kızlarım Merve ve Elif Erva 'ya sonsuz teşekkür ederim.

Uzun bir aradan sonra yüksek lisansa başlamam konusunda beni yüreklendirerek bu çalışmanın ortaya çıkmasında manevi katkıda bulunmuş ve yıllarca beraber çalıştığım bir ağabey olarak çok sevdiğim her zaman saygı duyduğum Sayın Dr. Ramazan YILMAZ'a çok teşekkür ederim.

Ocak 2014 Ramazan Pehlivan

İÇİNDEKİLER

KISALTMALAR

Bkz. : Bakınız
BM : Benzerlik Matrisi
CV : Cross Validation (Çapraz Doğrulama)
Dİ : Doküman İşleme
DTV : Dikey Tarama Vektörü
DVM : Destek Vektör Makinesi
Gİ : Görüntü İşleme
HTD : Horizontal Traverse Density (Yatay Çizgi Hareketliliği)
K-EYK : K-En Yakın Komşuluk
NB : Naive Bayes
OCR : Optical Character Recognition (Optik Karakter Tanıma)
RO : Rastgele Orman
SB : Satır Belirleme
SP : Satır Parçalama
VTD : Vertical Traverse Density (Dikey Çizgi Hareketliliği)
YTV : Yatay Tarama Vektörü
YSA : Yapay Sinir Ağı
yak. : Yaklaşık

TABLO LİSTESİ

ŞEKİL LİSTESİ

SİMGE LİSTESİ

d(i,j) : i. ve j. Birimin Birbirine Uzaklığı
$\mathbf{d_{k(i,j)}}$: k. Kümenin i. ve j. Kümelere Olan Uzaklığı
$\mathbf{d_{ki}}$: k. Kümenin i. Küme İle Olan Uzaklığı
$\mathbf{x_{ik}}$: i. Birimin k. Değişken Değeri
$\mathbf{x_{jk}}$: j. Birimin k. Değişken Değeri
P(A\B): B Olayının Gerçekleştiği Durumda A Olayının Meydana Gelme Olasılığı
P(A) : A Olayının Gerçekleşme Olasılığı

ÖZET

Üniversite	:	İstanbul Kültür Üniversitesi
Enstitüsü	:	Fen Bilimleri Enstitüsü
Dalı	:	Matematik-Bilgisayar
Programı	:	Matematik-Bilgisayar
Tez Danışmanı	:	Yard. Doç. Dr. Levent Çuhacı
Tez Türü ve Tarihi	:	Yüksek Lisans – Ocak 2014

RESİM TABANLI OSMANLICA BELGELERDE SINIFLANDIRMA

Ramazan Pehlivan

Bu çalışmanın amacı resim formatındaki Osmanlıca belgeleri içeriklerine göre sınıflandıran bir model ortaya koymaktır. Bu amaçla resim formatında taranmış Osmanlıca matbu belgelerde, ‘‘Görüntü İşleme’’, ‘‘Kümeleme’’ ve ‘‘Doğal Dil İşleme’’ tekniklerini birlikte kullanarak ‘‘Doküman Sınıflandırma’’ yapan etkin bir sınıflandırma yöntemi önerilmiştir.

Çalışmamızda veri olarak Türkiye Büyük Millet Meclisi (TBMM) Kütüphane ve Arşiv Hizmetleri Başkanlığı’nın resmi web sitesinden alınan Osmanlıca belge örnekleri seçilmiştir. Görüntü işleme teknikleriyle belgeler sayısal forma dönüştürülmüş, ardından satırlar ve satırlardaki kelime ya da harf grupları tespit edilmiş ve her bir harf grubu ayrı birer resim olarak kaydedilmiştir. Resimler arasında kümeleme yapılarak aynı (ya da benzer) harf grupları aynı kümeye atanmıştır. Harf gruplarının ait oldukları küme bilgileri kullanılarak bu belgelerin, birbirini izleyen etiket numaralarını içeren metin formatındaki karşılıkları elde edilmiştir. Bu aşamadan sonra doküman sınıflandırma alanında geçerli bir teknik olan kelime frekans analizi, elde ettiğimiz dönüştürülmüş metin dosyalarında küme frekans analizi olarak uygulanmıştır. Sonuç olarak; resim formatında taranmış Osmanlıca belgeler; semantik analize tabi tutulmadan, belgeyi oluşturan harf gruplarının benzerlik ölçütleri baz alınarak sınıflandırılmıştır.

Proje MATLAB ortamında geliştirilmiş ve bir makine öğrenmesi uygulaması olan WEKA programında sınıflandırma sonuçları elde edilmiştir. Ayrıca aynı veri seti üzerinde kelime frekans analizine dayalı bir doküman sınıflandırma uygulaması da gerçeklenmiştir.

Anahtar Kelimeler: Osmanlıca belge, doküman sınıflandırma, resim kümeleme, frekans analizi, satır parçalama, hiyerarşik kümeleme.

ABSTRACT

University	:	İstanbul Kültür University
Institute	:	Institute of Science
Department	:	Mathematic-Computer
Literature Programme	:	Mathematic-Computer
Supervisor	:	Assis. Prof. Dr. Levent Çuhacı
Degree Awarded and Date	:	MA – January 2014

CLASSIFICATION OF IMAGE-BASED OTTOMAN RECORDS

Ramazan Pehlivan

Aim of this work is developing a model which classifies image-formatted Ottoman records by their contents. For this purpose, an effective classification method, which conjunctively uses "Image Processing", "Clustering" and "Natural Language Processing" techniques, is proposed for image-formatted scans of Ottoman printed records.

In our work, Ottoman record samples from the official web page of Turkish Grand National Assembly (TBMM) Library and Documentation Center were used as data. Records were converted into digital form via image processing techniques, then words or letter groups in documents were detected and stored separately as individual pictures. By clustering between these pictures, identical (or similar) letter groups were registered to the same cluster. By using cluster information of letter groups, text-formatted counterparts, which include consecutive label numbers, were obtained for records. After that step, word frequency analysis, which is a valid technique in document classification, was used on converted text files as cluster frequency analysis. As a result, image-formatted scans of Ottoman records were classified based on similarity criteria of constituting letter groups, without using semantic analysis.

Project was developed on MATLAB environment and classification results were obtained by a machine learning application software, WEKA. Another classification method based on word frequency analysis was also implemented using the same data set.

Keywords: Ottoman record, document classification, image clustering, frequency analysis, line segmentation, hierarchical clustering

1.GİRİŞ

Bilişim çağını yaşayan modern dünyada, bilgi üretimi çok hızlı gerçekleştiğinden doğru bilgiye kısa zamanda ulaşma önemli bir sorun haline gelmiştir. Özellikle internet yoluyla elektronik ortamda bilgi ve belge üretimi çok hızlı gerçekleşmektedir. İhtiyaç duyulan bilgiye, sınırsız sayıda belge arasından kolayca erişebilme ihtiyacı, doküman sınıflandırma problemini ortaya çıkarmıştır.

‘‘Doküman sınıflandırmadaki amaç, bir dokümanın özelliklerine bakılarak önceden belirlenmiş belli sayıdaki kategorilerden hangisine dahil olacağını belirlemektir. Doküman sınıflandırma, Bilgi Alma (*Information Retrieval*), Bilgi Çıkarma (*Information Extraction*), Doküman İndeksleme, Doküman Filtreleme, otomatik olarak data elde etme ve web sayfalarını hiyerarşik olarak düzenleme gibi pek çok alanda önemli rol oynamaktadır. Doküman sınıflandırmadaki problemlerden biri, kime ait olduğu bilinmeyen veya yazarının kimliğinden şüphelenilen dokümanların yazarının tahmin edilmesi, bir diğer problem de dokümanın türünün veya yazarının cinsiyetinin belirlenmesidir’’[21].

Örneğin; web sitelerinde haber akışı takip etme, haberlerin sınıflandırılması, ulusal güvenlik, ticari güvenlik, e-posta sınıflandırma, spam filtreleme [26], posta yönlendirme, alıntı (intihal) takibi [14], arşiv tarama gibi alanlarda metin sınıflandırma teknikleri kullanılır. Bunlara ek olarak yazar tanıma [15,25], metin konusunu belirleme [43], metin yazarının cinsiyetini belirleme [15,25], gibi konular da doküman sınıflandırma sistemlerinin çalışma alanlarıdır.

Belge sınıflandırma çalışmalarını; dijital ortamda yazılmış (metin) belgelerde ve resim formatında taranmış belgeler de olmak üzere, iki farklı zeminde inceleyebiliriz. Dijital ortamda yazılı olan belgelerde yapılan çalışmalar, ‘Doğal Dil İşleme’ tekniklerini kullanarak bir sınıflandırıcı yöntem yardımıyla, belgenin türünü belirlemektedir [2,10,22]. Resim tabanlı belgelerde ise öncelikle görüntü işleme teknikleri ve Optik Karakter Tanıma (*Optical Character Recognition-OCR*) ile bilgisayara tanıtılmakta, ikinci adımda ise yine ‘Doğal Dil İşleme’ ve değişik sınıflandırıcı yöntemler yardımıyla kategori belirlenmektedir [9,32]. Resim tabanlı

belgelerde OCR kullanmadan sınıflandırma yapılan çalışmalar da bulunmaktadır [1,5].

Teknolojik gelişmelere paralel olarak dokümanların dijital ortama aktarılması ve bunların işlenmesi, gerekli bilgiye ulaşım konusunda zamandan tasarruf sağlamaktadır. Ayrıca, günümüze kadar raflarda matbu eserler olarak korunmuş kütüphane arşivleriyle, resmi ve özel kurum arşivleri de resim formatında taranarak dijital ortama aktarılmaktadır. Böylece hem zaman ve yerden tasarruf edilmekte hem de kıymetli bilgiler içeren eserlerin güvenliği ve kalıcılığı sağlanmaktadır.

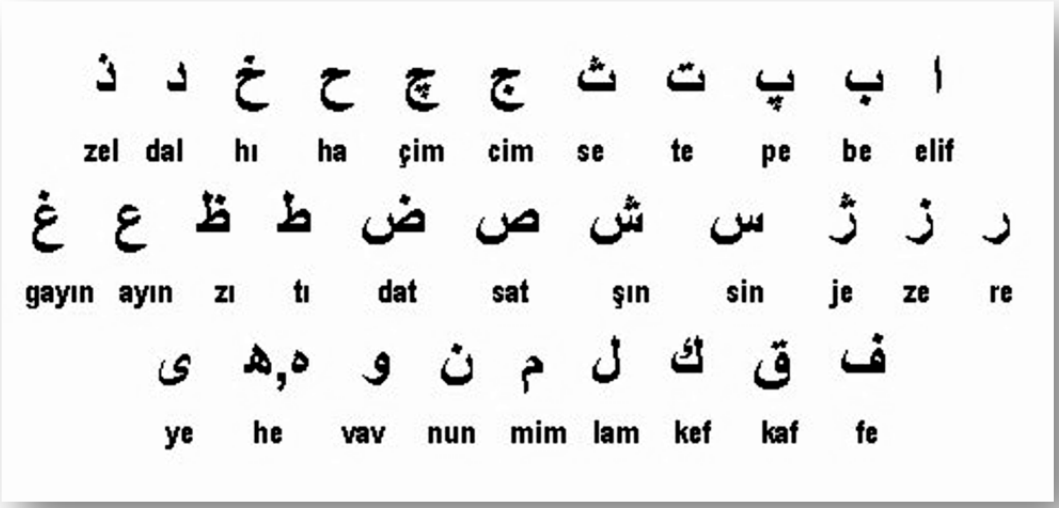
Ülkemizdeki devlet arşivlerinin büyük bir bölümü Osmanlıca ve Arapça eserlerden oluşmaktadır. Ancak, ‘‘Doküman sınıflandırma sistemlerinin çoğunluğu İngilizce yazılan dokümanları işlemek için tasarlanmıştır. Bu nedenle Arapça yazılan belgeler için geçerli değildir. Arapça metin sınıflandırma sistemleri geliştirme, Arapça’nın karmaşık ve zengin bir dil olması nedeniyle oldukça zordur’’ [2].

‘‘Ülkemiz Osmanlı İmparatorluğu'ndan devraldığı arşiv belgeleriyle, dünyanın en zengin arşivlerine sahip sayılı ülkelerinden biridir’’ [13]. Bu durumda Arapça ve Osmanlıca belgelerde yapılan doküman sınıflandırma işlemlerinin önemi daha iyi anlaşılmaktadır.

1.1 Osmanlıca’nın Yapısı

13-20. yüzyıllar arasında Anadolu’da ve Osmanlı Devleti’nin hüküm sürdüğüyerlerde yaygın olarak kullanılmış olan, özellikle 15. yüzyıldan sonra Arapça ve Farsça’nın etkisinde kalan Türk yazı dili, ‘Osmanlıca’ olarak adlandırılır. Osmanlı Türkçesi ya da eski yazı olarak da bilinen Osmanlıca Arapça, Farsça ve Türkçe’nin karışımıdır ve Arap alfabesiyle yazılır [30] .

Arapça’da 28 harf, Osmanlıca’da ise 32 harf vardır. Osmanlıca’da Arap harflerinin yanı sıra Farsçadaki ‘p’ (پ), ‘ç’ (چ), ve ‘j’ (ژ) harfleri de mevcuttur. Bu 31 harfin dışında Türkçe’deki ince ‘g’ ünsüzünü belirtmek için kef harfine bir çizgi ekleyerek gef (گ) harfi de kullanılmıştır. Osmanlıca da Arapça gibi sağdan sola doğru yazılır.



Şekil 1.1 Osmanlıca Alfabe [30]

Arapça ve Osmanlıca' yı Latin alfabesinden ayıran ve bilgisayar tarafından tanınmasını zorlaştıran bazı özellikler vardır. Bu özellikleri şöyle sıralayabiliriz.

1) Harfler çoğunlukla birbirine bitişik şekilde yazılır. Bir harf grubu bazen bir kelimeyi gösterir bazen de bir kelime bir kaç parçalı harf grubundan oluşur. Dolayısıyla kelimeleri birbirinden ayırmak ilk hamlede zordur.

2) Harflerin bulunduğu konuma göre (başta,ortada,sonda) farklı yazılışları vardır (Bkz.Tablo1.1).

3) Bir dokümanda harfler farklı kalınlıklar gösterebilmekte, hatta aynı harf bir dokümanın farklı yerlerinde değişik kalınlıklarda yazılabilmektedir.

4) Latin alfabesinde herbir karakterin yatay sırada birbirini takip etmesine karşılık Arapça'da bazı karakterlerin üst ya da alt uzantıları birbirlerinin düşeyde hizalarına gelebilmektedir.

5) Osmanlıca dokümanlarda genelde satırlar birbirlerinden ayrılmış olmakla birlikte bazen bir satırın alt uzantısı, alttaki satırın üst sınırını veya bir satırın üst uzantısı, üstteki satırın alt sınırını aşmaktadır. Bu yüzden satırları birbirinden ayıran net bir hat bulunmamaktadır.

Bu özellikler doküman işlemenin temelini oluşturan ‘‘Alan Parçalama’’ ve ‘‘Alan Etiketleme’’yi zorlaştırmakta bu da sistemin çalışmasında hata oranını yükseltmektedir.

‘‘Latin harflerinin matbaa çıktıları üzerinde elde edilen sonuçlar günümüzde yeterli düzeye ulaşmıştır. Fakat Çince, Japonca ve Osmanlıca karakterler üzerinde sorun hala devam etmektedir. Özellikle Osmanlıca ve Arapça karakterlerin tanıtılması, dilin yapısı ve yazım şekli göz önünde tutulduğunda oldukça zorlaşmaktadır. Hatta matbaada hazırlanmış bir Osmanlıca metnin tanınması, el yazısı ile yazılmış bir Latince metinden daha zor olabilmektedir’’ [29].

Tablo 1.1 Osmanlıca Harflerin Başta, Ortada ve Sonda Yazılışları [30]

İsimleri	Harfler	Sonda (sağdan bitişik)	Ortada (her iki taraftan bitişik)	Başta (soldan bitişik)	Karşılıkları
elif	ا ا	ا	ـا	ا	a, e
hemze	ء أ ا	ء ا	ـئـ	ئـ	’`(a, e, i, u, ü)
be	ب	ب	ـبـ	بـ	b
pe	پ	پ	ـپـ	پـ	p
te	ت	ت	ـتـ	تـ	t
se	ث	ث	ـثـ	ثـ	s
cim	ج	ج ج	ـجـ ـچـ	جـ	c
çim	چ	چ چ	ـچـ ـجـ	چـ	ç
ha	ح	ح ح	ـحـ ـهـ	حـ	h
hı	خ	خ خ	ـخـ ـکـ	خـ	h
dal	د	د	ـد	ـد	d
zel	ذ	ذ	ـذ	ـذ	z
re	ر	ر	ـر	ـر	r
ze	ز	ز	ـز	ـز	z
je	ژ	ژ	ـژ	ـژ	j
sin	س	س س	ـسـ	سـ	s
şın	ش	ش ش	ـشـ	شـ	ş
sat	ص	ص ص	ـصـ	صـ	s
dat	ض	ض ض	ـضـ	ضـ	d, z
tı	ط	ط ط	ـطـ	طـ	t
zı	ظ	ظ ظ	ـظـ	ظـ	z
aynn	ع	ع ع	ـعـ	عـ	‘ ,
gaynn	غ	غ غ	ـغـ	غـ	g, ğ (kalın)
fe	ف	ف ف	ـفـ	فـ	f
kaf	ق	ق ق	ـقـ	قـ	k
kef	ك	ك ك	ـكـ	كـ	k, g, ğ (y), n
gef	گ	گ گ	ـگـ	گـ	g, ğ
nef, sağır kef	ڭ	ڭ ڭ	ـڭـ	ڭـ	n
lam	ل	ل ل	ـلـ	لـ	l
mim	م	م م	ـمـ	مـ	m
nun	ن	ن ن	ـنـ	نـ	n
vav	و	و	ـو	ـو	v, o, ö, u, ü
he	ه	ه ه	ـهـ ـه	هـ	h, e, a
lamelif	لا	لا	ـلا	ـلا	la
ye	ى	ى	ـيـ	يـ	y, ı, i

1.2 Osmanlı Arşivleri

‘‘Ünlü Osmanlı tarihçisi Prof. Dr. Halil İNALCIK ‘Bana Osmanlı Arşivleri’ni verin size bir kültür imparatorluğu kurayım’ diyerek Osmanlı Arşivleri'nin önemini çok veciz bir şekilde ortaya koymuştur. Bilindiği üzere, her millet bir tarihî mirasın sahibidir. Bu tarihî mirasın çok önemli bir bölümünü arşivler, kütüphâneler ve eski eserler gibi maddî ve manevî kültür varlıkları teşkil ederler. Millet olabilme ve kalabilmede bu kültür varlıklarının büyük yeri ve rolü vardır. Arşivler, devletin ve fertlerin haklarını ve milletlerarası münasebetleri belgeler ve korurlar. Bir konuyu aydınlatmaya ve tespite yararlar. Bu arada ait olduğu devrin örf ve âdetlerini, sosyal yapısını, meselelerini ve bunlar arasındaki münasebetleri ortaya koyarlar. Türkiye, arşiv malzemesi bakımından çok büyük zenginliğe sahiptir. Osmanlı Devleti'nden devralınan büyük mirasla, bugün dünyanın en zengin arşiv potansiyeline sahip sayılı ülkelerden birisi durumundayız.

Bugün dünyada 19'u Arap, 11'i Balkan ve Avrupa, 3'ü Kafkas, 2'si Orta Asya Türk devleti, 2'si Kıbrıs, İsrail ve Türkiye Cumhuriyeti olmak üzere toplam 39 bağımsız ülke Osmanlı Devleti'nin hükümrân olduğu coğrafya üzerinde yer almaktadır. Bu ülkelerin Osmanlı dönemlerindeki tarihlerinin en zengin kaynağı Osmanlı Arşivleri'dir.

Orta ve Yakın Doğu, Balkan ve Akdeniz ülkeleri içerisinde kudretli ve kuvvetli devlet olabilme vasfını uzun süre devam ettiren Osmanlı Devleti'nde, arşiv fikri çok eskilere kadar uzanmaktadır. Arşivin, bir milletin tarih ve kültür hazinesi olduğunu idrâk eden ecdâdımız, bunun içindir ki, kurduğu arşiv teskilâtına "Hazîne-i Evrâk" adını vermiştir […] ’’ [8].

‘‘Osmanlı devlet belgeleri çok iyi tutulur, sağlam kâğıtlara, silinmez mürekkeple yazılır ve çok iyi muhafaza edilirlerdi. *Defter emini*, istenen defter ve vesikayı, milyonlarca defter ve vesika arasından birkaç dakika içinde bulabilirdi. Çünkü en iyi şekilde ve fevkalade tasnif edilmişlerdi[…]. Şu anda 100 milyonun üzerinde tarihi vesika bulunduran *Başbakanlık Osmanlı Arşivi* yalnız Türkiye'nin değil, Osmanlı İmparatorluğu'nun sona ermesinden sonra kurulan 40 a yakın devletin de ana arşivi durumundadır’’[31].

‘‘Arşivlerimizdeki belgeler, yalnızca Türkiye için değil Avrupa, Kuzey Afrika ve Yakındoğu ülkelerinin siyasi, iktisadi ve kültürel konulardaki sorunlarının çözümünde de önem taşımaktadır. Bu nedenle arşivlerimizde araştırma yapmak isteyen yerli ve yabancı araştırmacıların sayısı da gün geçtikçe artmaktadır’’[13].

Başbakanlık Osmanlı Arşivleri’ndeki belgeler halen taranarak elektronik ortama aktarılmaktadır. Arşivlerdeki belgeler 400 uzman tarafından elle tasnif edilmeye çalışılmaktadır [17]. Milyonlarca belgenin elle tasnifinin ne kadar zor olduğu ve ne kadar uzun süre gerektirdiği aşikardır. Osmanlı arşivlerinin ülkemiz ve dünya için önemi düşünüldüğünde, günümüzün teknolojik gelişmelerinden yararlanılarak, belgelerin en kısa sürede araştırmacıların hizmetine sunulması gerektiği anlaşılmaktadır.

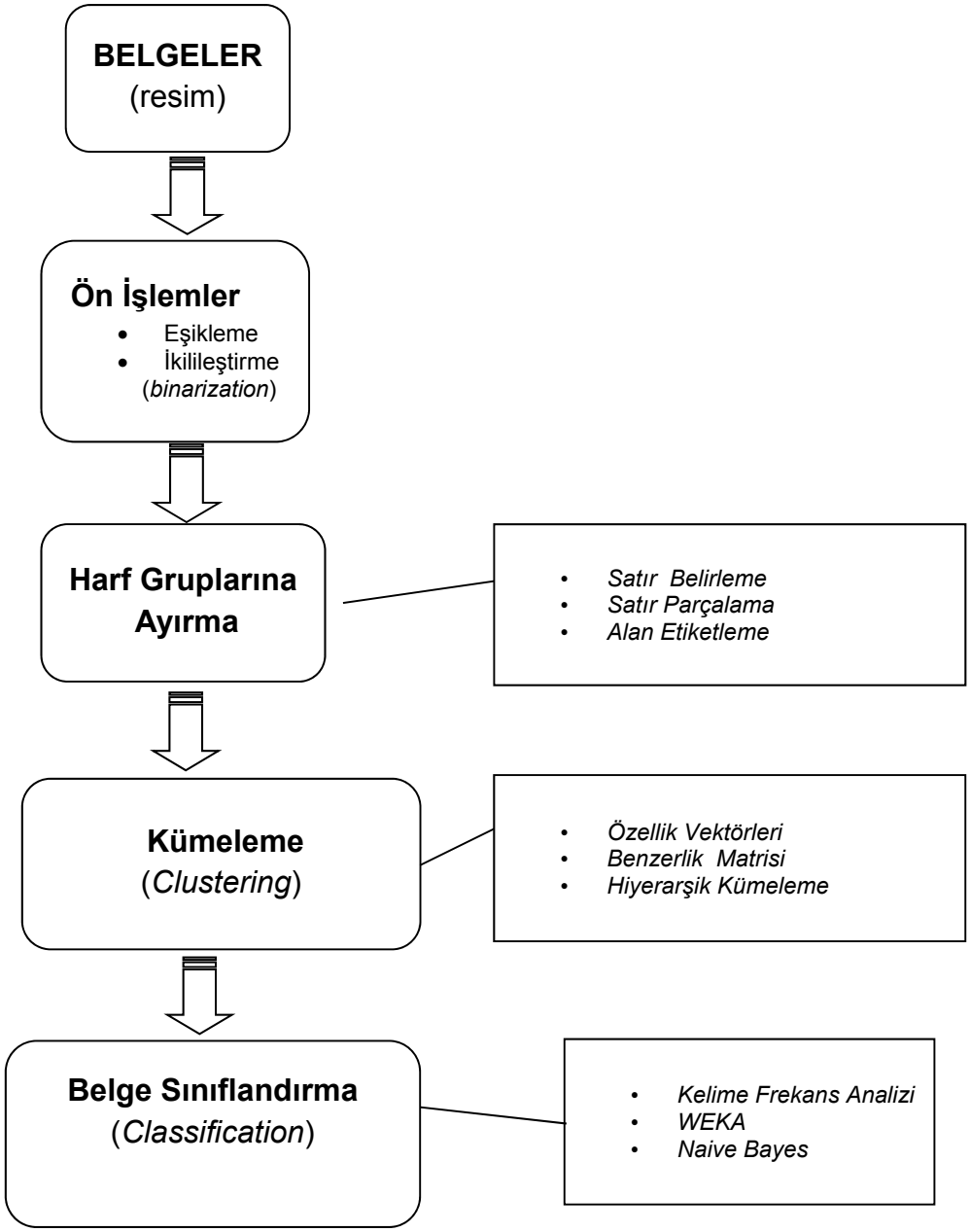
Osmanlı arşivlerindeki eserlerin bir kısmı matbu olup, büyük çoğunluğu el yazmasıdır. El yazısı karakterlerin kişiden kişiye farklılıklar göstermesi, harflerin birleşmesi, matbaa çıktılarına göre tam bir standardının olmaması ve özellikle Osmanlıca’daki çok çeşitli ve süslü yazı tipleri nedeniyle, el yazması eserlerin bilgisayar tarafından algılanması çok zordur. Bu alanda yapılan başarılı çalışmalar [2,3,30] olmasına rağmen henüz optimal bir sistem geliştirilememiştir.

Bu çalışmada, elektronik ortamda bulunan matbu haldeki Osmanlıca belgelerin, bilgisayar tarafından sınıflandırılabilmesi için yeni bir model ortaya konmuştur. Kullandığımız yöntemde ‘Görüntü İşleme’ (*Image Processing*) ve ‘Doğal Dil İşleme’ (*Naturel Language Processing*) teknikleri bir araya getirilerek resim formatında taranmış Osmanlıca arşiv belgelerinin, bilgisayar tarafından anlamsal (*Semantic*) analize girmeden sınıflandırılması amaçlanmıştır.

İlk olarak resim formatındaki taranmış belgeler, ikilileştirme (*Binarization*) ile sayısal (0-1) forma getirilmiştir. Bir sonraki aşamada belgedeki satırların alt ve üst sınırları piksel bazında tespit edilerek satırlar belirlenmiş ve satır parçalama ile satırları oluşturan harf grupları farklı birer resim olarak kaydedilmiştir. Her bir resmin özellik vektörleri çıkarılmış, özellik vektörleri ikili kombinasyonlar halinde karşılaştırılarak belirlenen benzerlik\farklılık puanları ile resimler arasında kümeleme (*Clustering*) yapılmıştır. Hiyerarşik kümeleme (*Hierarchical Clustering*) sonucunda benzer resimlere aynı etiket numarası verilmiştir. Böylelikle belge, harf gruplarının

sayısal etiket numaraları birer kelime gibi düşünülmek suretiyle; resim formatından ardışık dizilmiş sayısal karakterlerden oluşan özel bir metin formatına dönüştürülmüştür. Belgelerde, her bir sayısal etiket numarasının kaç kez geçtiğini içeren özellik vektörleri çıkarılmış ve bir makine öğrenmesi uygulaması olan WEKA paket programında sınıflandırma sonuçları elde edilmiştir. Ayrıca özel olarak yazılan bir programla, kelime frekans analizine göre sınıflandırma yapan bir çalışma gerçeklenmiştir. Böylece Doğal Dil İşleme'nin en yaygın yöntemlerinden olan ve karakterlerin dokümanda yer alma sıklığına göre sınıflandırma yapan kelime frekans analizi, tezimizde resim formatında taranmış Osmanlıca belgelere uygulanabilmiştir. Denemelerde, Türkiye Büyük Millet Meclisi (TBMM) Kütüphane ve Arşiv Hizmetleri Başkanlığı'nın resmi web sitesinden üç farklı sınıftan (roman, sosyoloji, tarih) 50'şer tane olmak üzere toplam 150 belge ile çalışılmıştır. Yapılan çalışmalarda MATLAB programından yararlanılmıştır.

Modelimizin genel adımlarını belirten blok diyagram Şekil 1.2 de görülebilir.



Şekil 1.2 Modelimizin Genel Adımlarını Belirten Blok Diyagram

2. DOKÜMAN SINIFLANDIRMA ALANINDA YAPILMIŞ ÇALIŞMALAR

Doküman sınıflandırma alanında ilk çalışmalar 70' li yıllarda karşımıza çıkmaktadır. Belli konularda özel sözlükler oluşturulmuş ve bu sözlük içindeki kelimeler birer kategori olarak atanarak doküman sınıflandırma yapılmıştır.

Doküman sınıflandırma alanındaki çalışmaları metin formatında yazılmış belgelerde ve resim formatında taranmış belgelerde olmak üzere iki başlıkta inceleyeceğiz.

2.1 Metin Formatında Yazılmış Belgelerde Yapılan Çalışmalar

Keselj ve arkadaşları (2003), yazar tanıması yaptıkları çalışmalarında değişik boyutlarda n- gram yöntemini kullanmışlar ve İngilizce, Yunanca ve Çince'ye uygulayarak karşılaştırmalı sonuçlarını vermişlerdir [33].

Diri ve Amasyalı (2003), yazar belirleme için Türkçe dokümanlar üzerinde yaptıkları çalışmada metin içeriğine bağlı sınıflamada Navie Bayes yöntemini, 22 farklı stil özelliğini kullanan diğer bir sınıflamada ise kendi geliştirdikleri '*Automatic Author Detection for Turkish Text'* metodunu kullanmışlardır'' [34].

Doğan ve Diri (2006), Türkçe bir dokümanın türü yazarı ve cinsiyetini belirlemek için üç farklı veri seti üzerinde yaptıkları çalışmada n-gram yöntemini kullanmışlardır. Naive Bayes (NB), Destek Vektör Makinesi (DVM), Rastgele Orman (RO), K-En Yakın Komşuluk (K-EYK) gibi sınıflandırıcıların yanında kendi geliştirdikleri *Ng-ind'* yöntemini de kullanarak testler yapmış ve başarı performanslarını birbirleri ile karşılaştırmışlardır. *'Ng-ind'* yönteminin cinsiyet ve tür belirlemede diğer yöntemlere göre daha iyi sonuçlar verdiğini gözlemlemişlerdir [15].

Khreisat (2006), Arapça metinlerin sınıflandırılması için çalışmıştır. Bunun için Arapça çevrimiçi gazetelerden derlediği metinlerde n-gram frekansı kullanarak ''Manhattan Benzerliği'' ve ''Dice'' benzerlik ölçütünü kullanmıştır [2].

Zaki ve arkadaşları (2010), Arapça belgelerdeki çalışmalarında; klasik Boole modelinden ilham alarak belgeleri oluşturan terimler arasındaki yakınlığı, fuzzy ilkesiyle anlamlandırarak sınıflandırma yapan bir model teklif etmişlerdir [12].

2.2 Resim Formatında Taranmış Belgelerde Yapılan Çalışmalar

Huang ve arkadaşları (2003), resim tabanlı belgelerde sınıflandırma için kelime şekil analizine dayalı bir yöntem teklif etmişlerdir. Bu çalışmada, OCR yerine doğrudan doğruya kelime resim özellikleri çıkarılır ve belge görüntüleri kelime birimleri halinde dik yönde parçalanır. Daha sonra dikey çubuk desenleri elde edilerek doküman özellik vektörleri oluşturulur. Son olarak doküman özellik vektörlerinin skaler çarpımlarından ikili benzerlik ölçütleri bulunarak sınıflandırma yapılır [4].

Özhan (2005), çalışmasında el yazısı ayrık yazılmış Osmanlıca harfleri tanımaya ilişkin bir yapay sinir ağı (YSA) tasarlamış ve uygulamıştır. Osmanlıca harflerin yazılı olduğu taranmış bir belge görüntü işleme teknikleri kullanılarak sayısal verilere dönüştürülmüştür. Verilerin düzenlenmesi için bir normalizasyon isleminden geçirilerek, YSA için giriş-çıkış değerleri elde edilmiştir.YSA'nın eğitim işlemi uygulamaları yapıldıktan sonra deneysel sonuçların değerlendirilmesi yapılmıştır [30].

Tan ve arkadaşları (2006), OCR kullanmadan görüntülü belgelere erişim için bir metot önermişlerdir. Belgeler karakter bölümlerine ayrılarak, her bir karakterin resim görüntüsü alınır. Her sütundaki dikey çizgi sayısı (Vertical Traverse Density - VTD) ile her satırdaki yatay çizgi sayısı (Horizontal Traverse Density -HTD) birer vektör şeklinde alınarak karakter resimlerinin görüntü özellikleri çıkarılmıştır. Bu özelliklere bağlı olarak bir n-gram tabanlı belge vektörü oluşturularak, belgeler arasında metin benzerliği, vektörlerin skaler çarpımı ile ölçülmüştür [1].

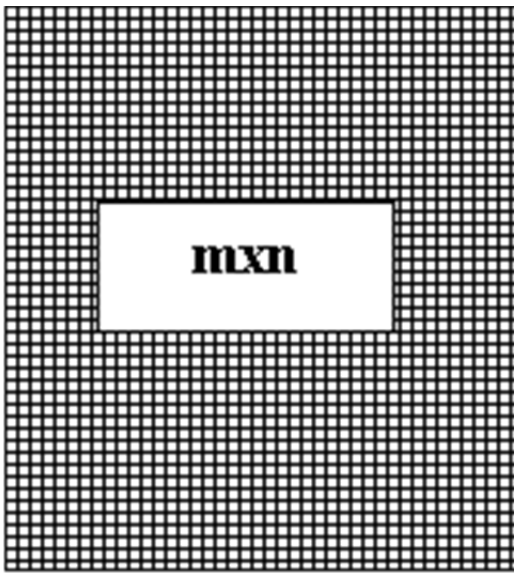
Yalnız ve arkadaşları (2009), resim formatında taranmış, basılı Osmanlı belgelerindeki çalışmalarında bağlı harfler için entegre edilmiş segmantasyon ve bir karakter tanıma modeli önermişlerdir. Önerilen model ilk olarak belli bir yazıdaki bir dizi segmenti ayıklar ve hangi bölümlerin en benzer olduklarını belirler. Daha sonraki işlem bu aday harflerin herbirinin sözdizimsel olarak doğru sırada puanlanmasıdır. Puan fonksiyonunu maksimize eden aday harfler sırasını bulmak

için, çevrimsiz yönetilen grafik oluşturulmuştur. Harfler bu grafikteki en uzun yolun hesaplanmasıyla tanınmıştır. Önerilen yöntem %90 doğruluk sağlamıştır [3].

3. GÖRÜNTÜ İŞLEME (IMAGE PROCESSING)

Dijital resimlerin, bilgisayar ortamına aktarıldıktan sonra görüntüden istenilen bilgilerin elde edilebilmesi ya da görüntü üzerinde istenilen değişikliklerin yapılabilmesi için uygulanan işlemlerin tümüne 'Görüntü İşleme' (Gİ) denir.
Resimlerin bilgisayar ortamında işlenebilmeleri için bilgisayar ortamına uygun hale getirilmeleri gerekmektedir. Bu dönüşüme sayısallaştırma (*digitizing*) adı verilir.

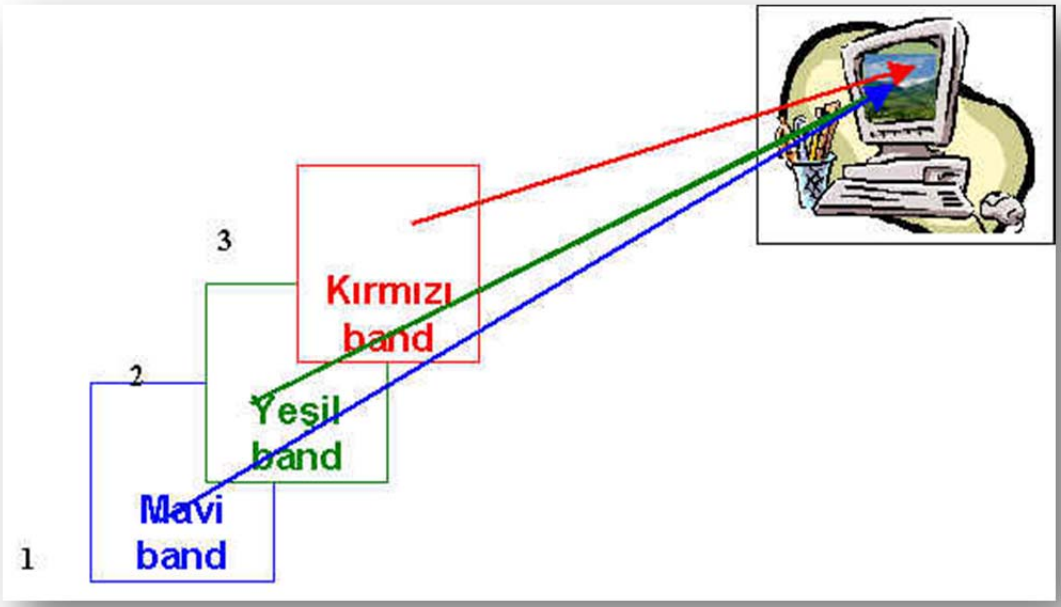
Bir görüntünün temel bileşeni pikseller (*pixel-picture element*) dir. Pikseller bireysel olarak üzerine düşen görüntüye ait renk ve parlaklık değerini taşıyan sayısal değerleri yansıtırlar. Resmin sözü edilen sayısal değerlerle ifadesine "sayısal görüntü" (*digital image*) denir. Sayısal görüntü satır ve sütünlardan oluşan bir matristir. Satır ve sütünların kesiştiği her bölgeye piksel denir. Dolayısı ile görüntü deyince piksellerdeki renk ve parlaklık değerlerinin saklandığı *mxn* boyutlu bir matris akla gelmelidir.



Şekil 3.1 Pikseller [38]

Yüksek çözünürlüklü ve renk ayrıntılarını taşıyacak biçimde algılanmış bir görüntü, daha küçük piksellerden oluşmakta yani piksel sayısı artmakta ve daha fazla sayısal veri ile temsil edilebilmektedir. Siyah beyaz bir resim iki boyutlu bir matris ile ifade edilirken renkli bir resim, Şekil 3.2 de görülebileceği gibi görüntünün renk tonlarını

oluşturan kırmızı-yeşil-mavi (*Red-Green-Blue, RGB*) kodlarını tutan 3 kademeli matris grubuyla ifade edilir [38].



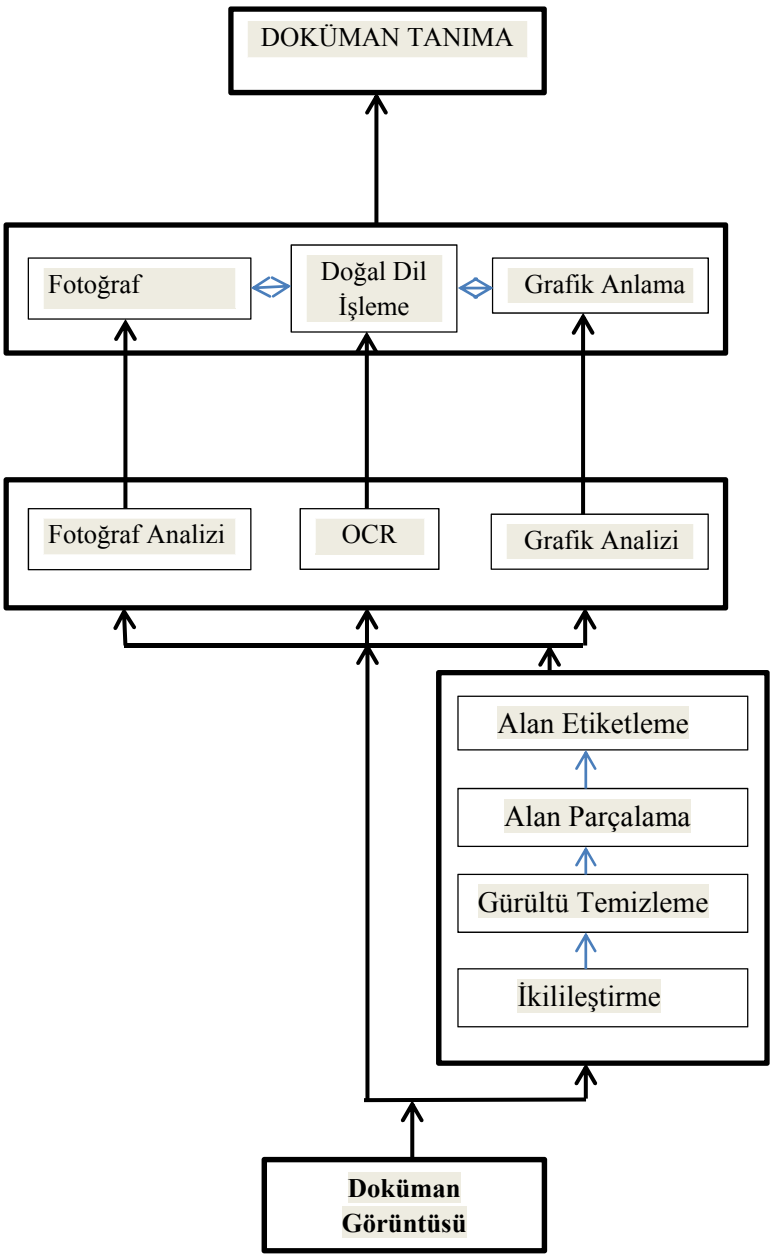
Şekil 3.2 Renkli Resimlerin Sayısal İfadesi [38]

4. RESİM FORMATINDA TARANMIŞ BELGELERDE GÖRÜNTÜ İŞLEME

Üzeri satırlarca yazı dolu olan bir kağıt okuma bilmeyen bir çocuk için bir şey ifade etmediği gibi, bilgisayarlar için de değersiz bir resimden ibarettir. Bilgisayarın bu resmin içindeki bilgiyi kullanabilmesi için yorumlaması gerekmektedir. Başka bir deyişle, bu resimdeki yazıların bilgisayarın anlayacağı daha kolay ve verimli olarak saklayıp bulabileceği sembollere çevrilmesi gerekmektedir [19]. Kısaca, resim formatındaki belge üzerinde yapılan Gİ işlemlerine 'Doküman İşleme' (Dİ) denmektedir. Dİ'nin önemli adımlarından biri belge üzerindeki nesne ve karakterleri tanıma teknikleridir. Bu işlemi gerçekleştiren sistemlere 'Optik Karakter Tanıma' (*Optical Character Recognition-OCR*) sistemleri denir.

Modern OCR sistemlerinin günümüzde kullanım alanı oldukça yaygındır. Banka çeklerinin, posta adreslerinin, anket formlarının okunması ve işlenmesi; havayolu bilet okuyucuları, sahte imza tespiti gibi insan eliyle yapılması çok zor olan birçok işlem bugün OCR sistemleri yardımıyla çok hızlı ve kolayca yapılabilmektedir. Özellikle el yazısının bilgisayar tarafından tanınması OCR' ın ilgilendiği en önemli problemdir. Bu alanda oldukça başarılı çalışmalar olmasına [27,29,32,45] rağmen hatasız çalışan bir sistem henüz geliştirilememiştir. Bununla birlikte günümüzde hızla yaygınlaşan tablet bilgisayarlarda ve akıllı cep telefonlarında elle ekran yüzeyinden girilen karakterlerin anlık olarak dijital sembollere dönüştürülmesi OCR tekniklerinin geldiği seviyeyi göstermektedir.

Genel bir Dİ sisteminin genel yapısı Şekil 4.1 deki gibidir.



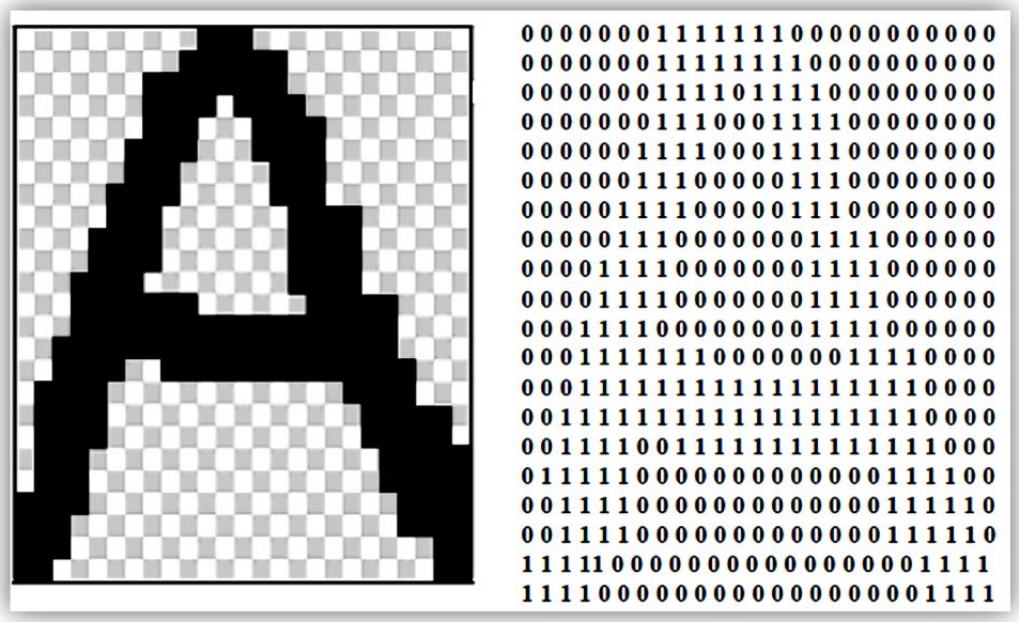
Şekil 4.1 Bir Doküman İşleme Sisteminin Genel Yapısı

Bu çalışmada Dİ sisteminin İkilileştirme, Alan Parçalama (Satır Belirleme-SB, Satır Parçalama-SP), Alan Etiketleme (Harf gruplarını etiketleme) adımları uygulanmıştır.

4.1 İkilileştirme (Binarization)

'Dİ' siteminde dokümanın renkli resmi yerine siyah beyaz görünümü üzerine çalışmak çoğu zaman yeterli olmaktadır. Eldeki görüntü siyah ile beyaz arası gri tonlardan oluşur veya renk özellikleri ilgili gri tonlara dönüştürülür. Resmi oluşturan pikseller, (8 bitlik piksel tanımlama formatında) farklı seviyedeki gri tonları, 0-255 arasında seviye değerleri ile temsil ederler. Burada 0 (siyah) dan 255 (beyaz) e kadar her piksel kendi gri tonunu temsil eden seviye değerini alır.

Resmi oluşturan siyah ve beyaz arası 256 farklı gri tonu işlem kolaylığı açısından, belli bir eşik değerine göre indirgenerek 0 veya 1 ile temsil edilir. Resim karanlık ise eşik değeri, orta seviye olan 128 değerinden yüksek; resim aydınlık ise 128 den düşük seçilir. Bu indirgemeye göre resimleri ikili hale getirmeye 'İkilileştirme' (*Binarization*) denir. Burada 1 siyah pikselleri, 0 ise beyaz pikselleri temsil eder (Bazı uygulamalarda ise 0 siyah pikselleri, 1 beyaz pikselleri temsil edebilmektedir). Bu aşamada iki renkli (siyah-beyaz) bir resim elde edilmiş olur.



Şekil 4.2 Bir 'A' Harfi ve İkili (*binary*) Görüntüsü [27]

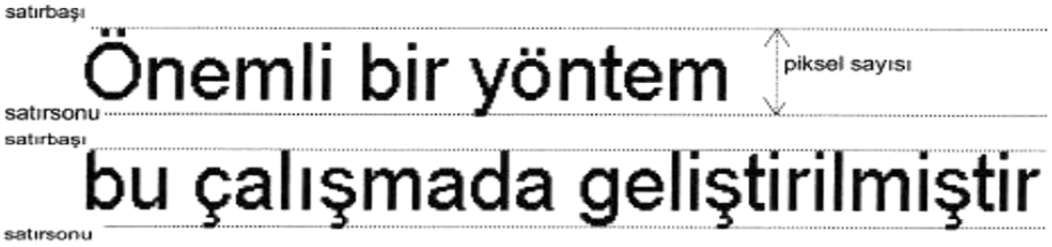
4.2 Gürültü Temizleme

Taranmış görüntüde, arka zeminde bulunan fakat asıl resme ait olmayan istenmeyen lekeler bulunabilir. Bunlara Gİ terminolojisinde ‘gürültü’ adı verilir. Sayısal forma geçme aşamasında gürültülerin yok edilmesi gerekir. Aksi takdirde pikseller üzerindeki farklı lekeler görüntünün bir parçası gibi algılanır bu da resmin tanımlanmasında hatalara neden olur. Görüntüdeki lekeleri yok etme işlemlerine ‘Gürültü Temizleme’ denir.

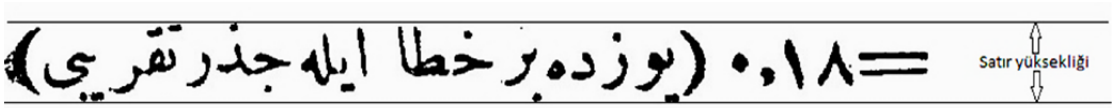
4.3 Satırların Belirlenmesi

Belgenin içeriğine ulaşabilmek için öncelikle tek bir resim halinde mevcut olan belgenin satırlarının tespit edilmesi gerekir. Bu işlemlere ‘‘Satır Belirleme- SB’’ diyoruz. SB işlemleri şu şekilde yapılır. Doküman görüntüsünün dikey başlangıç noktasından başlanarak her bir piksel satırı (görüntü matrisinin satırları), siyah pikselle karşılaşıncaya kadar yatay olarak taranır (Görüntü matrisinde ilk kez 1 rakamıyla karşılaşılan satıra kadar tarama yapılır). İlk siyah piksele rastlanılan piksel satırı, metnin başlangıç satırı olarak kabul edilir. Böylece metnin başlangıç satırının üst sınırı tespit edilmiş olur. Tarama işlemine devam edilerek siyah piksel içermeyen ilk satır (matriste sadece 0 lardan oluşan satır) bulunduğunda metnin başlangıç satırının bitimine (alt sınıra) ulaşılmış olur. Matrisin, üst sınırı veren satır numarası ile alt sınırı veren satır numarası arasındaki fark ilgili satırın yüksekliğidir (Bkz. Şekil 4.3 ve Şekil 4.4). Satır belirlemenin en önemli problemi; bazı karakterlerin alt ve üst noktalarının bulunduğu piksel satırlarının ayrı birer metin satırı gibi algılanmasıdır. Bunu önlemek için tespit edilen her satırın yüksekliği ile alt ve üst satırlara olan uzaklıklarına bakılır. Şöyle ki;

- İki satır arasındaki uzaklığın ikinci satıra oranı 0,2 yada daha küçükse,
- İlk satırın yüksekliğinin ikinci satırın yüksekliğine oranı 0,2 veya daha küçükse, iki satır birleştirilerek tek bir satır haline getirilir [19] .



Şekil 4.3 Latin Harfleriyle Örnek Satır Belirleme [19]

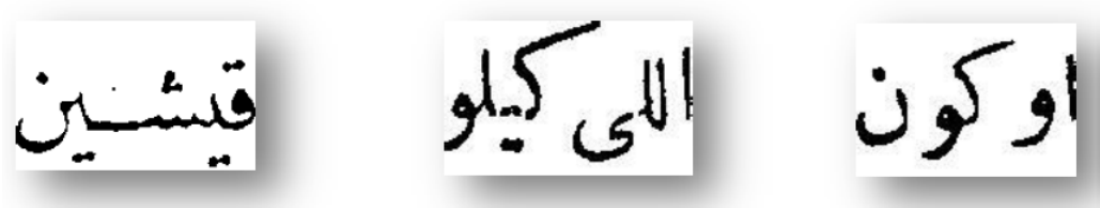


Şekil 4.4 Osmanlıca Yazıda Örnek Satır Belirleme

4.4 Satır Parçalama (Kelime/Harf Gruplarının Tespiti)

Satır belirleme işlemleri tamamlandıktan sonra 'Satır Parçalama-SP' işlemleri dediğimiz, her bir metin satırındaki kelimelerin tespit edilmesi gerekmektedir. Bu aşamada ilk olarak satırın üst ve alt sınırları arasında yatay başlangıç noktasından başlanarak dikey doğrultuda tarama yapılır. Osmanlıca'da sağdan sola, Latin alfabesinde ise soldan sağa tarama yapılır. SB'de olduğu gibi sadece beyaz piksellerden oluşan sütunlar geçilir, siyah piksel içeren ilk sütun, karakterin başlangıcı olarak kabul edilir. Tarama işlemine devam edilerek tamamı beyaz piksellerden oluşan ilk sütun ise kelime/harf grubunun sonu kabul edilir. Bu işlem tüm satırlara, satır başından sonuna kadar uygulanarak belgedeki kelimelere ulaşılmış olur. Latin alfabesi ile yazılı dokümanlarda karakter aralıkları ve kelime aralıkları standart olduğundan SP ile kelimelerin elde edilmesi kolaydır. Ancak Osmanlıca yazılarda karakter boşlukları standart değildir. Harfler genelde bitişik yazıldığından karakterlerin ayrı ayrı parçalanması zordur (Bkz. Şekil 4.5 (a)). Diğer taraftan bazı kelimeler ise bitişik yazılmayan harfler nedeniyle iki veya daha fazla parçadan oluşabilmektedir (Bkz. Şekil 4.5 (c)). Ayrıca bazı durumlarda dikey yönde uzantısı birbirinin üstüne gelen harflerden dolayı beyaz piksel sütununa

rastlanmadığından kelime boşluğu tespit edilemez ve parçalama gerçekleşmez. Böylece iki farklı kelime, iki resim gibi algılanması gerekirken, tek bir resim gibi işlem görebilir (Bkz. Şekil 4.5 (b)). Bu yüzden Latin alfabesinden farklı olarak, Osmanlıca'da karakteri tespit etmek yerine kelime ya da harf grubunu tespit etmek, anlamsal çıkarımın gerekli olmadığı durumlarda daha doğru bir işlem olacaktır.



(a) **(b)** **(c)**

Şekil 4.5 Taranmış Osmanlıca Belgeden Elde Edilen Örnek Harf Grupları

Elde edilen harf gruplarının çevresel boşluklardan arındırılması gerekir. Bunun için görüntünün resim matrisi, alt üst ve yanlardan taranarak tamamı 1'lerden oluşan (yani tamamen beyaz piksel içeren) satır ve sütunlar hariç tutularak; siyah piksel içeren ilk satır ve sütunun numaraları belirlenip bir alt matris oluşturulur. Bu alt matris, resmin çevresel boşluklarından arındırılmış şeklinin sayısal formudur.

4.5 Alan Etiketleme (Harf Gruplarının Etiketlenmesi)

SP işleminden sonra 'Alan Etiketleme' yapılır. Bu aşama satır parçalama işleminden sonra elde edilen kelime ya da karakter\harf gruplarına birer etiket ismi ya da etiket numarası verilme işlemidir.

5. BENZERLİK MATRİSİ

Belgelerde karakter veya harf gruplarına ulaşıldıktan sonra karakterlerin bilgisayar tarafından tanınması ve belgenin sınıflandırılabilmesi için optik karakter tanıma (*OCR*) ya da benzer nesneleri gruplandırmak için kümeleme (*clustering*) işlemlerine geçilir.

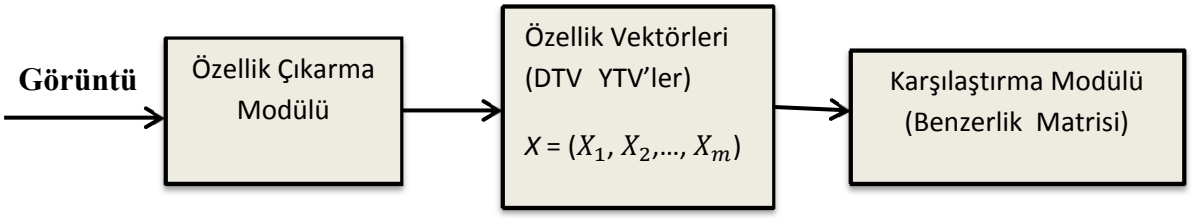
Kümelemede ilk aşama nesnelerin birbirlerine olan benzerlik ya da farklılıklarının hesaplanarak bir matrise alınmasıdır. Bu matrise ‘‘Benzerlik (veya Farklılık) Matrisi (BM)’’ adı verilir. Öncelikle uygun bir özellik çıkarma yöntemiyle resimlerin belirleyici özellikleri bir vektöre aktarılır ve bu vektörler üzerinden resimler arası benzerlik puanları hesaplanır.

Özellik çıkarma tekniklerinden biri aşağıda anlatılmıştır.

5.1 Özellik Çıkarma *(Feature Extraction)*

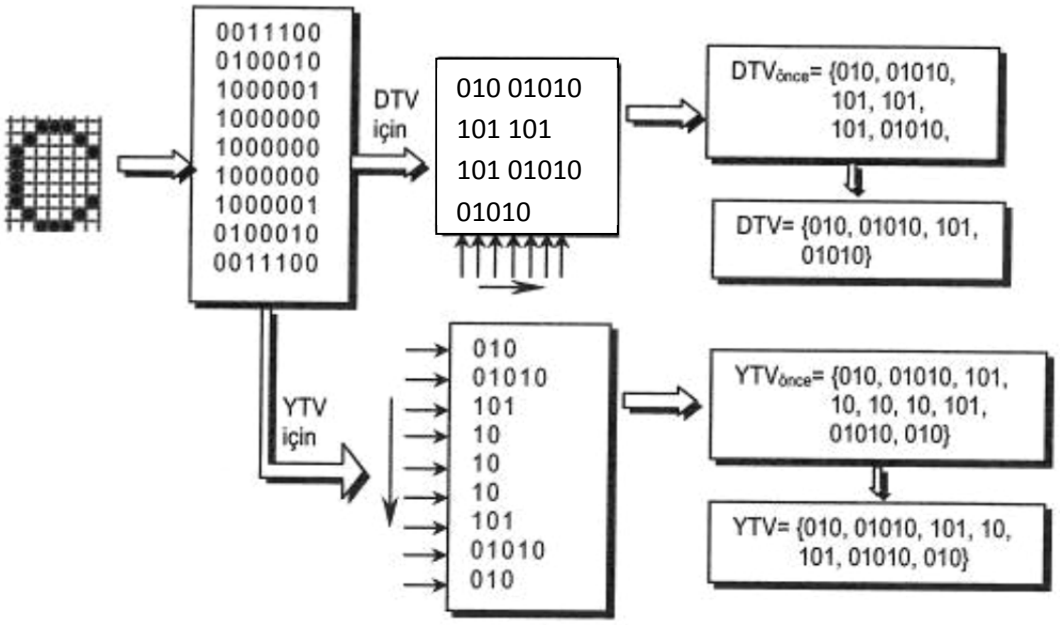
Özellik çıkarma; görüntü üzerinde bazı ölçümler yaparak, görüntünün özelliklerini çıkarmak ve bunları bir özellik vektörüne aktarmaktır.

SP den gelen görüntü dikey ve yatay olarak taranıp, görüntüye ilişkin Dikey Tanımlama Vektörü (DTV) ve Yatay Tanımlama Vektörü (YTV) elde edilir. Bu vektörler karakterlerin dikey ve yatay yönlerde sıkıştırılmış şeklidirler. Resmin boyut özellikleri ortadan kalktığından, aynı karakterin farklı boyutlardaki görüntülerinin özellik vektörleri birbirine benzerdir.



Şekil 5.1 Benzerlik Matrisinin Aşamaları

DTV' yi elde etmek için görüntünün her sütunu sol alt köşeden başlamak üzere dikey olarak taranır. Bu işlem sonunda örnek 'C' (Bkz.Şekil 5.2) karakterinin ilk sütunu için '001111100' dizisi bulunur. Bu dizinin ardışık 0 ve 1' leri tekleştirmek suretiyle '010' dizisi elde edilir. Tüm sütunlar için bu işlem tekrarlandığında $DTV_{\text{önce}}$ bulunur. Bu vektörde ardışık birbirine eşit vektör elemanlarının tekleştirilmesi ile de DTV elde edilir. Aynı işlemler yatay doğrultuda tekrarlandığında ise YTV elde edilmiş olur [19].



Şekil 5.2 'C' Karakteri İçin DTV ve YTV' nin Elde Edilişi [19]

Bu yönteme benzer olarak Tan ve arkadaşları (2006), görüntüdeki siyah-beyaz geçişleri yerine yatay ve düşey her sıradaki siyah çizgi sayılarına (hareketliliğine) göre özellik çıkaran bir yöntem önermişlerdir [1]. Bu yöntemde SP den gelen karakterin resim görüntüsünde, her sütundaki dikey çizgi hareketliliği (*Vertical Traverse Density* - VTD) ile her satırdaki yatay çizgi hareketliliği (*Horizontal Traverse Density* - HTD) birer vektör şeklinde alınarak karakter resimlerinin görüntü özellikleri çıkarılır.

Şekil 5.3 "S" ve "V'' Karakterleri İçin HTD ve VTD' nin Elde Edilişi [1]

Örnekte "S" ve "V'' karakterleri için, görüntünün satırları üst satırdan başlayarak aşağıya doğru her satır yatay şekilde taranır. Satırlardaki doğru parçası sayıları sırasıyla alınarak Şekil 5.3'deki gibi HTD vektörü oluşturulur. Benzer şekilde, VTD soldan sağa dikey tarama ile elde edilen başka bir vektördür. Şekil 5.3'de yatay ve dikey hat kesimleri, ilgili HTD ve VTD altında küçük kareler olarak temsil edilmiştir.

6. KÜMELEME (*CLUSTERING*)

Herhangi bir veri setini belli yöntemlerle analiz ederek daha önceden isimleri belli olmayan gruplara ya da alt gruplara ayırma işlemine 'Kümeleme (*Clustering*)' denir. Kümelemede en önemli nokta, küme içi benzerliklerin olabildiğince yüksek, kümeler arasındaki benzerliğin ise olabildiğince düşük olmasıdır.

Kümeleme analizi ilk kez 1939 yılında Tryon tarafından kullanılmıstır. 1960'lı yıllardan sonra kullanımı yaygınlaşmıştır. 1963 yılında Robert Sokal ve Peter Sneath'ın yazdığı "*Sayısal Sınıflandırma İlminin Temelleri*" adlı kitap bu alanda önemli bir adım olmuştur [40].

Kümeleme işleminin adımları aşağıdaki gibidir.

1) Dağınık yapıda bulunan veri setlerinden alınan ve hangi gruba ait olduğu bilinmeyen nesnelere ait özelliklerin çıkarılması.

2) Uygun bir benzerlik ölçüsü ile değişkenlerin birbirlerine olan uzaklıklarının hesaplanması (Benzerlik ya da farklılık matrisinin oluşturulması).

3) Benzerlik matrisindeki bilgilere göre, tercih edilen bir kümeleme yöntemiyle verilerin uygun sayıda kümelere ayrılması.

6.1 Kümeleme Analizinde Benzerlik\Uzaklık Ölçüleri

Nesneleri kümelerken, aralarındaki benzerliğin bulunması için değişik uzaklık ölçüleri kullanılır. Bu ölçülerden bazıları: Öklid (*Euclidean Distance*), Pearson, Manhattan, Minkowski uzaklık ölçüleridir. Bu çalışmada kullanılan ölçü ''Öklid'' uzaklık ölçüsüdür.

6.1.1 Öklid Uzaklık (*Euclidean Distance*) Ölçüsü

Öklid uzaklığı en sık kullanılan uzaklık ölçüsüdür.

Öklid uzaklık ölçüsü kullanılarak, p boyuta sahip iki birim arasındaki uzaklık :

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

formülüyle hesaplanır.

x_{ik} : i. birimin k. değişken değeri .

x_{jk} : j. birimin k. değişken değeri .

$d(i,j)$: i. ve j. birimin birbirine olan uzaklığı .

i =1, . . . , n ; j =1, . . . , n ve k =1, . . . , p 'dir. (n birim ve p değişken sayısıdır).[39]

6.2 Kümeleme Yöntemleri

Kümeleme yöntemleri hiyerarşik olmayan kümeleme yöntemleri (bölmeli yöntemler) ve hiyerarşik kümeleme yöntemleri olarak iki kategoride incelenebilir.

6.2.1 Bölmeli Yöntemler (Hiyerarşik Olmayan Yöntemler)

Bu metotlar, n adet birimden oluşan veri setini başlangıçta belirlenen k<n olmak üzere 'k' adet kümeye ayırmak için kullanılır. Bölmeli yöntemler arasında en yaygın olanı ''K-Ortalamalar'' (*K-Means*) yöntemidir.

6.2.1.1 K-Ortalamalar Kümeleme Yöntemi (*K-Means*)

K − Ortalama tekniğinde, işlemler şu sıra ile yapılır:

1. İlk olarak başlangıç küme merkezleri rastgele 'k' adet olarak seçilir.
2. Birimler, küme merkezlerine olan uzaklıklarına göre en yakın kümelere atanır. Aynı gruptaki değişkenlerin ortalamaları alınarak yeni küme merkezleri oluşturulur.
3. Değişkenlerin en yakın oldukları küme merkezlerine atamaları tekrar yapılır.

4. Bu işlemler yeni küme merkezlerine göre küme elemanlarının yerleri değişmez olana kadar tekrarlanır.

6.2.2 Hiyerarşik Kümeleme (*Hierarchical Clustering*) Yöntemleri

Hiyerarşik yöntem, benzerlik matrisini kullanarak birbirine en yakın (uzaklık değerleri en küçük) nesneleri birleştirmeye dayanmaktadır. Başlangıçta her nesne ya da gözlem ayrı bir küme olarak kabul edilir. Daha sonra birbirine en yakın iki küme yada gözlem birleştirilerek yeni bir küme elde edilir. Böylece küme sayısı bir azaltılmış olur. Yeni kümelerden birbirine en yakın olanlar tekrar birleştirilerek her adımda küme sayısı azaltılmış olur. Bu işlemler tekrarlanarak tek bir küme elde edinceye kadar devam edilir. Hiyerarşik kümeleme teorik olarak tek bir küme kalıncaya kadar devam etmesine rağmen bu istenen bir durum değildir. Kümeleme işleminin nerede durdurulacağı önemlidir. İstenilen küme sayısına ulaşınca ya da kümeler arasındaki uzaklık önceden belirlenmiş bir eşik değerini aşınca kümeleme işlemi sonlandırılır [28].

Bu teknikte eğer i ve j nci birimler birleştirilmiş ise birleştirilen kümenin k'nıncı küme ile ilişkisi uzaklık ölçütü olarak,

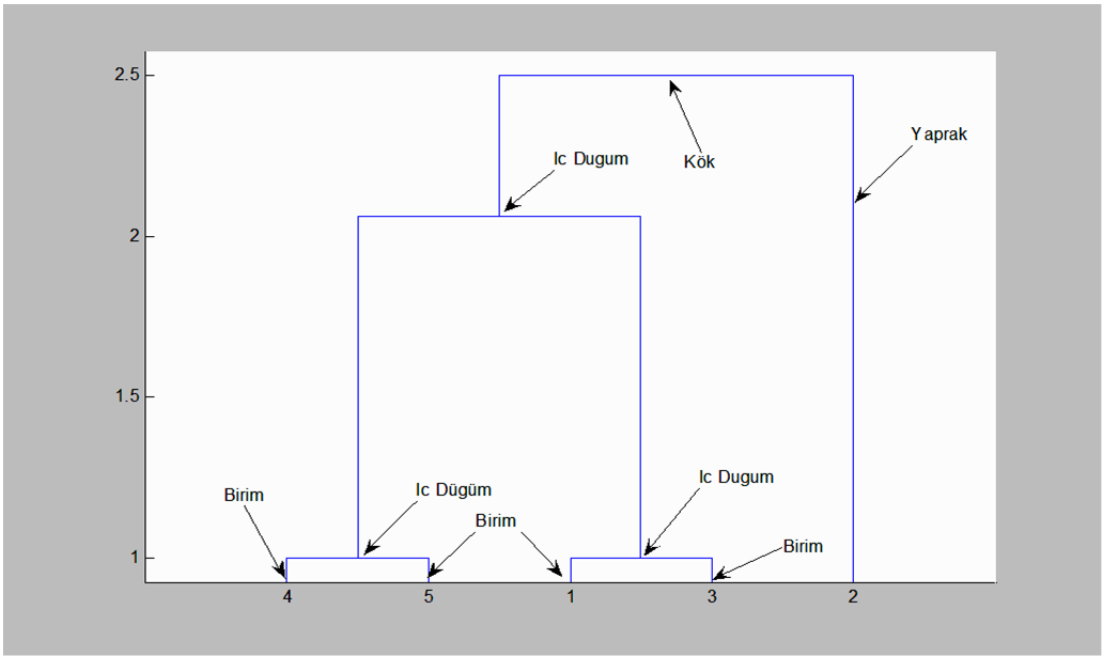
$$d_{k(i,j)} = \min(d_{ki}, d_{kj})$$

biçiminde ifade edilmektedir.

Eşitlikte; $d_{k(i,j)}$: k'nıncı kümenin daha önce oluşan i. ve j. kümelere olan uzaklığını, d_{ki} : k' nıncı kümenin i'nci küme ile olan uzaklığını,

d_{kj} : k' nıncı kümenin j'nci küme ile olan uzaklığını göstermektedir [28].

Hiyerarşik ayrıştırma sırasında, "ağaç veri yapısı" olarak da bilinen dendrogram kullanılır. Dendrogram, hiyerarşik kümeleme tekniğiyle elde edilen kümelerin görselleştirilmesini sağlar. Şekil 6.1 de örnek bir dendrogram yapısı görülmektedir.



Şekil 6.1 Örnek Dendrogram [39]

Benzerlik Matrisini oluşturan elemanların kümelenmesi için üç farklı uzaklık birimi kullanılır. Bir elemanın bir önceki adımlarda oluşturulmuş yeni kümeye olan uzaklığı hesaplanırken, kümedeki elemanların en yakınına göre birleştiren yöntem tek bağlantı (*single linkage*), kümedeki elemanların en uzağına göre birleştiren yöntem tam bağlantı (*complete linkage*), kümedeki elemanların uzaklık ortalamasına göre birleştiren yöntem ise ortalama bağlantı (*average linkage*) adı verilir.

Bu çalışmada ortalama bağlantı yöntemi kullanılmıştır.

6.2.2.1 Ortalama Bağlantı (Average Linkage)

Uzaklıklardan ya da benzerliklerden oluşan N*N kare matriste minimum uzaklıkta olan kümelerin birleştirilmesiyle oluşturulan yeni kümenin diğer kümelere olan uzaklıkları, yeni oluşturulan kümedeki elemanların ağırlık merkezi ile diğer kümelerin elemanlarının ağırlık merkezleri arasındaki uzaklıklardır. Elde edilen yeni matriste ise, birbirine en yakın olan kümeler birleştirilir.

7. DOKÜMAN SINIFLANDIRMA

Dünya bilgi çağına girdiğinden beri gelişen ülkelerde insanlar tarafından kullanılan bilginin miktarı çok hızlı bir şekilde artmaktadır. Bu bilgileri sağlıklı bir şekilde kullanabilmek ve kısa sürede erişebilmek için birbirleri ile ilişkili olan bilgileri bulup aynı bilgi topluluğu içinde toplamak gerekir. Bu da dokümanları sınıflandırmayı gerektirir [21]. Belge sınıflandırmada amaç, sınıfı belli olmayan bir dizi belgenin önceden bilinen gruplardan hangisine dahil olacağını belirlemektir.

Belge sınıflandırmada ilk olarak belgenin özellik vektörlerinin çıkarılarak her bir dokümanın özel bir şekilde elde edilmesi gerekir. Özellik çıkarımında kullanılan ‘‘N-gram’’ ve ‘‘Terim Frekans’’ istatistikleri en yaygın yöntemlerdendir

7.1 N-Gram Model

N-gram, bir karakter (harf) bloğunun ‘n’ adet karakterden oluşan dilimidir. N-gram tabanlı sınıflandırma yöntemi, doküman içerisindeki karakter tabanlı n-gram’ların kullanım frekansına dayalı bir işlemdir [15]. **‘**Aynı sınıf ya da kategoriye ait dokümanların benzer n-gram dağılımları vardır’ düşüncesinden çıkmıştır. İlk kez Damashek tarafından elektronik metinlerin benzerliğini ölçmek için önerilmiştir [1]. Dilden ve anlamdan bağımsızdır. Seçilen ‘N’ değerine göre ‘2-gram’, ‘3-gram’, ‘4-gram’ … vs. şeklinde kullanılır. Karakter n-gram veya kelime n-gram olarak da uygulanabilmektedir.

Karakter n-gram için, örnek cümlemiz; ‘‘ *Belge Sınıflandırma*’’ ise :

2-gramlar : ‘Be’, ‘el’, ‘lg’, ‘ge’, ‘e_’, ‘_S’, ‘Sı’, ‘ın’, ‘nı’, ‘ıf’, ‘fl’, ‘la’, ‘an’, ‘nd’, ‘dı’, ‘ır’, ‘rm’, ‘ma’

3-gramlar : ‘Bel’, ‘elg’, ‘lge’, ‘ge_’, ‘e_S’, ‘_Sı’, ‘Sın’, ‘ını’, ‘nıf’, ‘ıfl’, ‘fla’, ‘lan’, ‘and’, ‘ndı’, ‘dır’, ‘ırm’, ‘rma’

4-gramlar : ‘Belg’, ‘elge’, ‘lge_’, ‘ge_S’, ‘e_Sı’, ‘ _Sın’, ‘Sını’, ‘ınıf’, ‘nıfl’, ‘ıfla’ ‘flan’, ‘land’, ‘andı’, ‘ndır’, ‘dırm’, ‘ırma’

Kelime n-gram için örnek cümlemiz ‘‘*Bu çalışma görüntü işleme ve belge sınıflandırma üzerine yapılmıştır*’’ ise:

kelime 2-gramlar : ‘Buçalışma’, ‘çalışmagörüntü’, ‘görüntüişleme’, ‘işlemeve’, ‘vebelge’, ‘belgesınıflandırma’, ‘sınıflandırmaüzerine’, ‘üzerineyapılmıştır’

kelime 3-gramlar :‘Buçalışmagörüntü’, ‘çalışmagörüntüişleme’, ‘görüntüişlemeve’, ‘işlemevebelge’,‘vebelgesınıflandırma’,‘belgesınıflandırmaüzerine’, ‘sınıflandırmaüzerineyapılmıştır’

Eğitim kategorisini oluşturan veri setinin ve sınıflandırılacak her bir dokümanın n-gram frekans istatistikleri çıkarılır. Yani her n-gram’ın ilgili belgede ve ilgili veri setinde kaç kez geçtiğinin istatistiği tutulur. Dokümanda bulunan n-gramlar en yüksek frekanstan en düşük frekansa doğru sıralanarak özellik vektörleri elde edilir. Daha sonra eldeki özellik vektörleri uygun bir sınıflandırıcıya verilerek belge sınıflandırılır.

7.2 Terim Frekansları

Bu temsil yönteminde metinler içerdikleri terimlerin frekanslarıyla ifade edilir. Bu terimler kelimelerin kendileri, kökleri ya da karakter gramlar olarak belirlenebilir. Bu yönteme göre satırlarında metinlerin, sütunlarında terimlerin yer aldığı bir matris oluşturulur. Matrisin [i, j] gözünde i. metinde j. kelimenin kaç kere geçtiği bilgisi tutulur. Matrisin satır sayısı metin sayısına, sütun sayısı ise tüm metinlerde geçen farklı kelimelerin sayısına eşittir [45] (Bkz.Şekil 7.1).

$$\begin{array}{c|cccc} & T_1 & T_2 & \dots & T_t \\ D_1 & d_{11} & d_{12} & \dots & d_{1t} \\ D_2 & d_{21} & d_{22} & \dots & d_{2t} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & d_{n1} & d_{n2} & \dots & d_{nt} \end{array}$$

$D_1, D_2, \dots D_n$: *Dokümanlar*

$T_1, T_2, \dots T_t$: *Dokümanlarda geçen farklı terimler*

Şekil 7.1 Örnek Terim Frekans Matrisi

Örnek: Metnin kelime frekanslarıyla ifadesi

‘‘Kardemir Karabük, Beşiktaş karşısında 2-0 galip, Fenerbahçe Kasımpaşa 2-2 berabere ve Trabzonspor, Çaykur Rize karşısında 3-2 galip’’.

Tablo 7.1 Metnin Kelime Frekans ve Oran Olarak İfadesi

Kelimeler	**Frekans**	**Oran**
Kardemir	1	0.043
Karabük	1	0.043
Beşiktaş	1	0.043
karşısında	2	0.086
2	4	0,174
-	3	0.13
0	1	0.043
galip	2	0.086
Fenerbahçe	1	0.043
Kasımpaşa	1	0.043
berabere	1	0.043
ve	1	0.043
Trabzonspor	1	0.043
Çaykur	1	0.043
Rize	1	0.043
3	1	0.043

Yukarıdaki tabloda, bir belgeyi temsilen verilen örnek metinin kelime frekansları ve kelime geçiş sayıların belgedeki toplam kelime sayısına oranı sırasıyla ‘‘ V’’ ve ‘‘V_{oran}’’ vektörlerine aktarıldığında, belgeyi tanımlayan özellik vektörleri aşağıdaki gibi oluşur.

$V = (1,1,1,2,4,3,1,2,1,1,1,1,1,1,1)$

$V_{oran=}(0.043,0.043,0.043,0.086,0.174,0.13,0.043,0.086,0.043,0.043,0.043,0.043,0.043,0.043,0.043,0.043)$

7.3 Sınıflandırma Yöntemleri

7.3.1 Naive Bayes (NB)

Naïve Bayes Sınıflandırma teoremi adını İngiliz matematikçi Thomas Bayes'ten (yak. 1701 - 1761) alır [18]. Doküman sınıflandırmada çok sık kullanılan ve diğer sınıflandırıcılara göre daha doğru sonuçlar veren bir yöntemdir. Dokümanın ait olduğu sınıf, kelimelerin ve sınıfların birleşik olasılıkları kullanılarak belirlenir.

Bayes teoreminin genel ifadesi (1) deki gibidir.

$$P(A/B) = \frac{P(B/A).P(A)}{P(B)} \tag{1}$$

P(*A*|*B*) : B olayı gerçekleştiği durumda A olayının meydana gelme olasılığıdır.
P(*B*|*A*) : A olayı gerçekleştiği durumda B olayının meydana gelme olasılığıdır.

P(A) ve **P(B) :** A ve B olaylarının olasılıklarıdır [18].

Naive Bayes teoremini doküman sınıflandırma problemine uygulayalım. Elimizde n adet sınıf olduğunu farz edelim $(S_1, S_2, \ldots, S_n)$. Herhangi bir sınıfa ait olmayan bir veri örneği X, verilen sınıflara ait olma olasılığı en yüksek değere sahip olan sınıfa atanır. Sonuç olarak, Naive Bayes sınıflandırıcı bilinmeyen örnek X' i, S_i sınıfına atar. Her veri örneği X, m boyutlu özellik vektörleri ile gösterilir.

$$X = (X_1, X_2, \ldots, X_m)$$

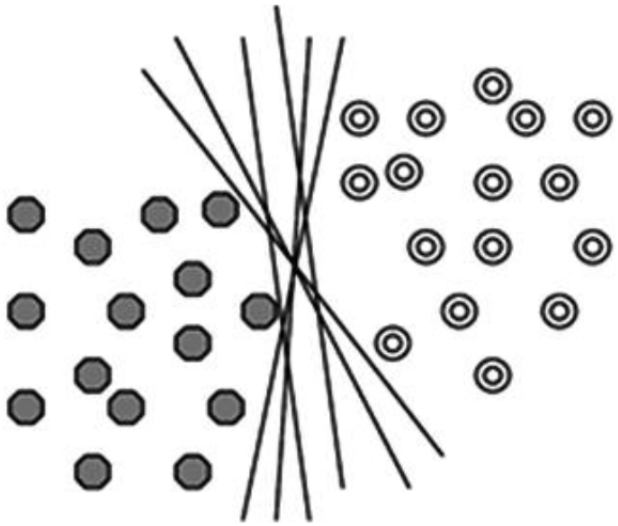
Özelliklerin hepsi aynı derecede önemlidir ve birbirinden bağımsızdır. Bir özelliğin değeri başka bir özellik değeri hakkında bilgi içermez. X örneğinin S_i sınıfında olma olasılığı (2) deki gibidir [15].

$$P(S_i/X) = \frac{P(X/S_i).P(S_i)}{P(X)} \tag{2}$$

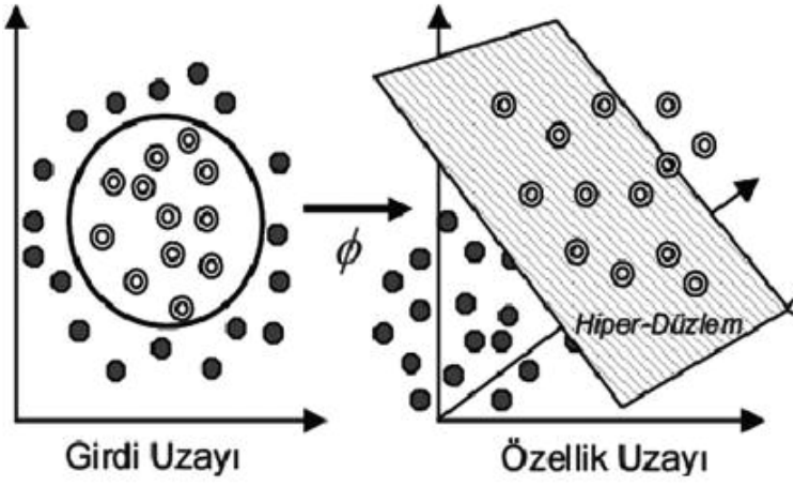
7.3.2 Destek Vektör Makinesi (DVM)

DVM 1960 lı yılların sonunda Vladimir Vapnik ve Alxey Chervonenkis tarafından geliştirilmiş istatistiksel tabanlı bir makine öğrenmesi yöntemidir. Son yıllarda özellikle veri madenciliğinde değişkenler arası örüntünün bilinmediği veri setlerindeki sınıflandırma problemi için kullanılır. Temel olarak iki boyutlu problemlerin çözümü için düşünülmüş, daha sonra çok boyutlu ve doğrusal olarak ayrılamayan problemlerin çözümüne de genelleştirilmiştir. DVM eğitim esnasında gözlemlenmemiş yeni verileri de sorunsuz olarak sınıflandırabilmektedir. Bu yönüyle diğer sınıflandırıcı yöntemlere göre iyi bir alternatif olmaktadır.

DVM doğrusal olarak ayrılabilecek verileri ayırabilen sonsuz sayıdaki doğrudan marjini en yüksek yapacak doğruyu seçmeyi hedefler. Doğrusal olarak ayrılamayan veriler doğrusal olarak ayrılabildikleri daha yüksek boyutlu başka bir uzaya taşınır ve bu uzayda optimum çalışan bir hiper düzlem bulmaya çalışır [44].



Şekil 7.2 Doğrusal Olarak Ayrılabilen İki Sınıflı Sınıflandırma Problemi [44]



Şekil 7.3 Doğrusal Olarak Ayrılamayan Verilerin Yüksek Boyutlu Uzaylara Taşınması [44]

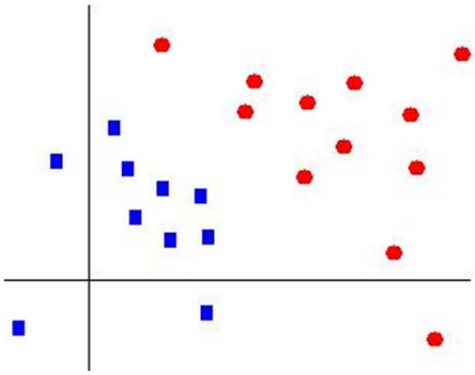
7.3.3 K-En Yakın Komşuluk (K-EYK)

Dokümanları özellik uzayındaki en yakın 'k' sayıda örneklerine göre sınıflandıran bir danışmanlı öğrenme tekniğidir. Nesneler arasındaki uzaklık hesabı için genellikle öklid uzaklığı kullanılır.

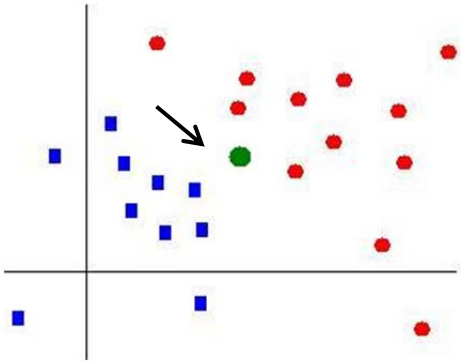
Sınıflandırılmak istenen örneğin sınıfı belirlenirken eğitim kümesinde o örneğe en yakın olan k adet örnek seçilir. Seçilen örnekler en çok hangi sınıfa ait ise ilgili test örneği de o sınıfa dahil edilir [21].

Bu metodun en büyük avantajı basit olmasıdır. Aynı zamanda gürültülü veriye karşı dirençlidir ve eğitim dokümanlarının sayısı fazla ise daha iyi sonuç verir. K-en yakın komşuluğun dezavantajı sınıflandırma için harcanan sürenin ortalamanın üzerinde olmasıdır. Bu sürenin uzun olmasının sebebi olarak herhangi bir ön hazırlık veya öğrenme fazı uygulanmaması söylenebilir [21]. Diğer dezavantajları ise şöyle sıralanabilir: 'k' parametresine ihtiyaç duyar, en iyi sonucun alınabilmesi için hangi uzaklık ölçümünün uygulanacağı ve hangi özelliklerin alınacağı bilgisi açık değildir, tüm dokümanlar vektörel olarak temsil edilir ve sorgu dokümanı ile diğer dokümanlar arasındaki kosinüs benzerliği hesaplanır.

Örneğin k = 3 için yeni bir eleman sınıflandırılmak istensin. Bu durumda eski sınıflandırılmış elemanlardan en yakın 3 tanesi alınır. Bu elamanlar hangi sınıfa dahilse, yeni eleman da o sınıfa dahil edilir. Mesafe hesabından genelde öklit mesafesi (*euclid distance*) kullanılabilir. Aşağıda verilen ve özelliklerine göre 2 boyutlu koordinat sistemine yerleştirilmiş olan örnekleri (Şekil 7.4 (a)) ele alalım.

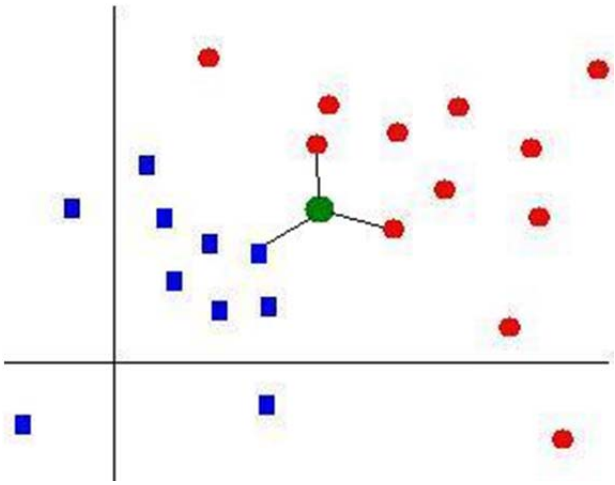


Şekil 7.4 (a)



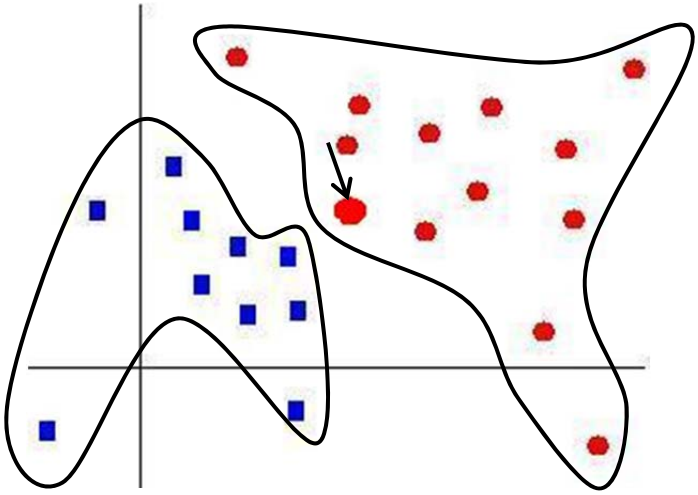
Şekil 7.4 (b)

K-EYK yöntemine göre Şekil 7.4 (b) de yeni bir üyenin geldiğini düşünelim ve Şekil 7.4 (c) deki gibi bu yeni gelen üyenin en yakın olduğu 3 komşu üyeyi tespit edelim.



Şekil 7.4 (c)

En yakın 3 üyenin iki tanesi yuvarlak üyeler olduğuna göre yeni üyemizi Şekil 7.4 (d) deki gibi sınıflandırabiliriz.



Şekil 7.4 (d) K-EYK Örnek Sınıflandırma [37]

Bunların dışında Rastgele Orman (*Random Forest*), Ögrenmeli Vektör Kuantalama (*Learning Vector Quantization*), Regresyon (*Regression*), Yapay Sinir Ağları (*Artificial Neural Networks*) ve Karar Ağaçları (*Decision Trees*) gibi teknikler de sınıflandırma uygulamalarında sıklıkla kullanım alanı bulunmaktadır.

8. UYGULAMA VE MATERYAL

Bu çalışmada materyal olarak, Türkiye Büyük Millet Meclisi (TBMM) Kütüphane ve Arşiv Hizmetleri Başkanlığı'nın resmi web sitesinden [42] resim formatında taranmış üç farklı sınıftan (roman, sosyoloji, tarih) 50'şer tane olmak üzere toplam 150 belge ile çalışılmıştır. Her bir veri setinin 35'er sayfası eğitim, 15'er sayfası ise test verisi olarak kullanılmıştır.

Modelin geliştirilmesi için yazılım materyali olarak iki hazır programdan yararlanılmıştır.

8.1 Kullanılan Yazılımlar

Kod yazımı ve uygulamaların çalıştırılması için ‘‘MATLAB’’ paket programı, belgelerin sınıflandırılması için ise bir makine öğrenmesi yazılımı olan ‘‘WEKA’’ programı kullanılmıştır.

8.1.1 Matlab

‘‘MATLAB’’ programı (Matrix ve Laboratory kelimelerinin birleşimiyle isimlendirilmiştir) matematiksel tabanlı bir sayısal hesaplama ve mühendislik yazılım paketidir. MATLAB matris işlenmesine, fonksiyon uygulama ve çizimlerine, algoritmalar uygulanmasına, kullanıcı arayüzü oluşturulmasına ve diğer dillerle yazılmış programlar ile etkileşim oluşturulmasına izin verir. C, C++, Java, ve Fortran dillerini içerir [36].

Her türlü grafiksel sonucun alınmasına izin verdiği için kullanım alanı çok geniştir. Özellikle doğrusal cebir, sayısal analiz öğretiminde ve görüntü işleme alanında çalışan bilim adamları arasında popülerdir. Ayrıca istatistik, mühendislik, ve ekonomi gibi alanlarda olduğu kadar endüstriyel işletmelerde de yaygın olarak kullanılmaktadır.

8.1.2 Weka

Weka, makine öğrenimi amacıyla 1993 yılında, ‘‘*University Of Waikato*’’ tarafından geliştirilmiş ve "*Waikato Environment for Knowledge Analysis*" kelimelerinin baş harflerinden oluşmuş yazılımın ismidir. Günümüzde yaygın kullanımı olan çoğu makine öğrenimi algoritmalarını içermektedir. Java dilinde geliştirilmiş olması ve kütüphanelerinin ‘jar’ dosyaları halinde geliyor olması sayesinde, JAVA dilinde yazılan projelere kolayca entegre edilebilmesi kullanımını daha da yaygınlaştırmıştır. Weka, tamamen modüler bir tasarıma sahip olup, içerdiği özelliklerle veri kümeleri üzerinde görselleştirme, veri analizi, iş zekası uygulamaları, veri madenciliği gibi işlemler yapabilmektedir [36]. Veri ön işlemesi (*data preprocessing*), regresyon (*regression*), sınıflandırma (*classification*), kümeleme (*clustering*), özellik seçimi veya özellik çıkarımı (*feature extraction*) da Weka’ nın yapabildiği işlemlerden bazılarıdır. Ayrıca bu işlemler sonucunda çıkan neticelerinde görsel olarak gösterilmesini sağlayan görüntüleme (*visualization*) araçları bulunmaktadır [37]. Weka yazılımı, kendisine özgü olarak bir ‘.arff ’ uzantısı ile gelmektedir. Temel olarak aşağıdaki temel işlemler Weka ile yapılabilir:

1. Sınıflandırma (*Classification*)
2. Kümeleme (*Clustering*)
3. İlişkilendirme (*Association*)

Ayrıca Weka kütüphanesinde veri kümelerini içeren dosyalar üzerinde çalışan çok sayıda hazır fonksiyon bulunmaktadır [36]. Bu çalışmada ‘Belge Sınıflandırma’ aşamasında Weka programından yararlanılmıştır.

8.1.2.1 ‘Arff’ Dosya Yapısı

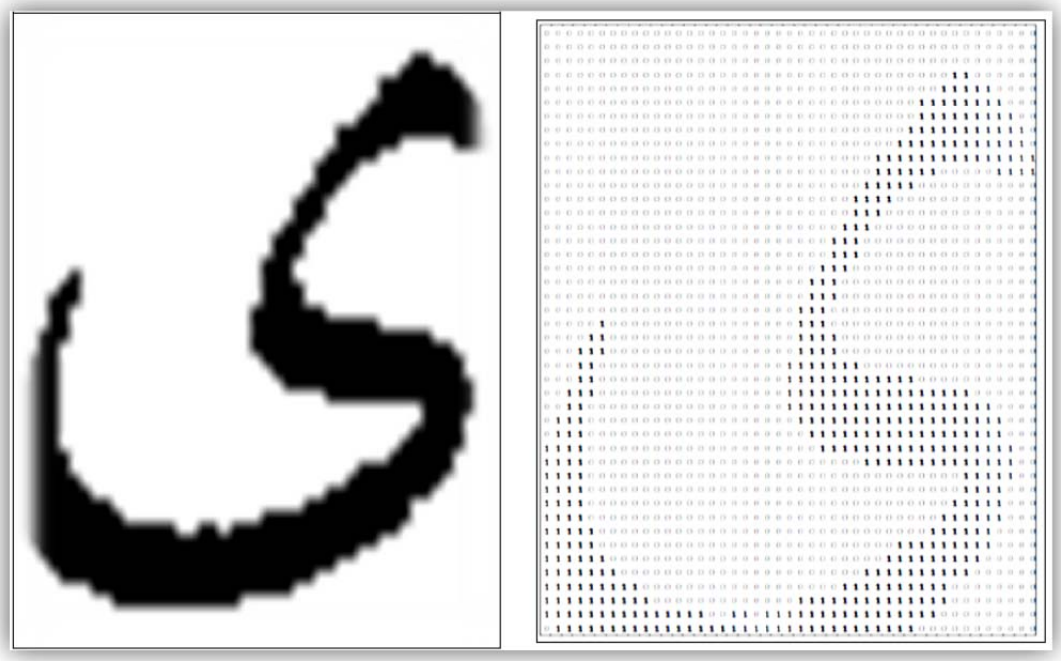
İngilizce, ‘‘*Attribute Relationship File Format*’’ kelimelerinin baş harflerinden oluşmuştur. ARFF dosya yapısı, Weka'ya özel olarak geliştirilmiştir ve dosya, metin yapısında tutulmaktadır. Dosyanın ilk satırında, dosyadaki ilişki tip (*relation*) tutulmakta olup ikinci satırdan itibaren de veri kümesindeki özellikler (*attributes*) yazılmaktadır. Özelliklerin hemen ardından veri kümesi yer alır ve veri kümesindeki her satır bir örneğe (*instance*) işaret etmektedir. Ayrıca veri kümesindeki her örneğin her özelliği arasında da virgül ayracı kullanılmaktadır.

8.2 Uygulama

İlk olarak resim formatındaki taranmış belgeler, ikilileştirme (*Binarization*) ile sayısal (0-1) forma getirilmiştir.

8.2.1 İkilileştirme

İkilileştirme için MATLAB' ın hazır '*im2bw*' fonksiyonu kullanılmış, eşik değeri olarak orta seviye seçilmiştir. Bu fonksiyonun çıktısı, resmin siyah noktaları için 0, beyaz noktaları için 1 değerlerini içeren iki boyutlu bir matristir.



Şekil 8.1 Osmanlıca 'y' Harfi ve İkili (*Binary*) Matrisi

8.2.2 Satır Belirleme (SB)

Eski Türkçe belgelerde satır belirleme için Latin alfabeyle yazılan belgelerde kullanılan yöntemler geçerlidir. Fakat Latin alfabesinde kısmen karşılaşılan satır aralarındaki karakter alt veya üst noktalarını içeren piksel satırlarının ayrı bir satır olarak algılanması, Eski Türkçe ile yazılan belgelerde (Osmanlıca'nın karakter özelliklerinden dolayı) daha büyük bir sorun olarak karşımıza çıkmaktadır. İkilileştirme adımından gelen matrise uygulanan işlemin algoritması aşağıdaki gibidir.

Algoritmanın girdileri :

I(m,n) : *mxn* boyutlarındaki resim tabanlı bir dokümana ilişkin ikili (*binary*) matris.

th : Dokümanda tespit edilen satırların yüksekliklerinin en büyük satır yüksekliğine oranını belirleyen alt eşik değeri.

Algoritmanın yerel değişkenleri:

hist(m): Resimdeki piksel satırlarına ilişkin siyah piksel sayılarının tutulduğu dizi.

Algoritmanın çıktısı :

r(k,2) : Dokümandaki 'k' adet gerçek satıra ilişkin piksel satırı başlangıç–bitiş indisleri matrisi.

1. **I** matrisini oluşturan her piksel satırına ilişkin siyah piksel sayılarını ***hist*** dizisinde sakla.

2. ***hist*** dizisinde başı ve sonu *0* ile belirlenmiş olan sayı bloklarının başlangıç ve bitiş indislerini ***r*** matrisinde sakla (bu bloklar siyah piksellere sahip piksel satırları olup, dokümandaki gerçek satırları belirlemektedir).

3. 2. adımda oluşturulan ***r*** matrisini taramak suretiyle en yüksek gerçek satırın piksel sayısını hesapla (***r*** matrisinde **r(i,2)- r(i,1)** değeri, dokümanın *i.* satır yüksekliğinin satır piksel sayısı cinsinden değerini verir).

4. ***r*** matrisindeki her satır için;

a. Satır yüksekliği, en büyük satır yüksekliğinin ***th*** katından daha büyük ise bu satırı geçerli bir satır olarak işaretle.

b. Satır yüksekliği, en büyük satır yüksekliğinin ***th*** katından daha küçük ise, bu satırı geçersiz bir satır olarak işaretle.

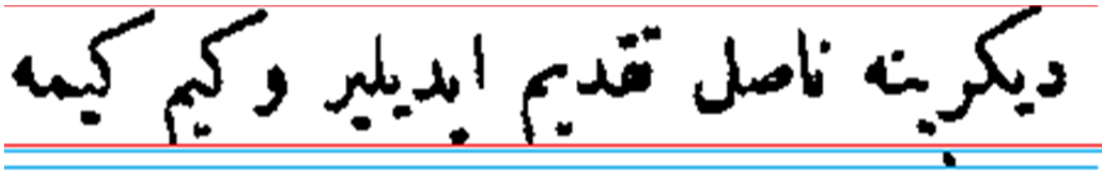
5. 4. adımda belirlenen her geçersiz satır için;

a. Geçersiz satır, bir üstteki geçerli satıra, alttaki geçerli satırdan daha yakın ise bu satırı üstteki satır ile birleştir (üstteki geçerli satırın alt sınırını genişlet).

b. Geçersiz satır, bir alttaki geçerli satıra daha yakın ise bu satırı alttaki satır ile birleştir (alttaki geçerli satırın üst sınırını genişlet).

Eski Türkçe alfabesindeki bazı karakterlerin alt ve üstlerine konan noktalama işaretlerinden dolayı satır tarama işleminde siyah satır bloklarından bazılarının yükseklik değerleri küçük çıkmaktadır. Algoritmanın belirlediği satırların gerçek birer satır olması; dokümandaki en yüksek satırın satır yüksekliğinin belirli bir eşik değeri (uygulamada 0.2 olarak seçilmiştir) ile çarpılması sonucu çıkan değer ile yapılan bir karşılaştırma sonucu belirlenmektedir. Algoritmanın 4. adımında yapılan bu karşılaştırma sonucunda yüksekliği çok küçük çıkan geçersiz satırlar, kendisine en yakın satırın bir uzantısı olarak düşünülüp bu satıra eklenmektedir (5. adım).

Örnek olarak Şekil.8.2'deki satır için, algoritma tarafından iki farklı satır belirlemesi yapılmıştır. Kırmızı çizgiler arasında belirlenmiş olan ana (gerçek) satır, her biri sıfırdan farklı sayıda siyah piksel içeren piksel satırlarından oluşur. Bu satıra ilişkin son piksel satırının ardından, tamamen beyaz piksel içeren az sayıda satır yer almakta; bu satırların da altında mavi çizgi ile belirlenmiş olan ve yüksekliği çok küçük olan bir başka satır belirlenmiştir. Mavi çizgi ile belirlenen satırın yüksekliği, dokümandaki en yüksek satırın eşik değeri (*th*) ile çarpılması sonucu çıkan değerden küçük olacağından algoritma tarafından geçersiz satır olarak işaretlenmiş ve bu satır kendisine en yakın geçerli satır ile birleştirilmişir.



Şekil 8.2 Satırların Belirlenmesi

piksel_satır_no	***hist dizisi***	
0	0	
1	**6**	
2	**10**	
3	**18**	
4	**40**	
..	..	Gerçek satıra ilişkin siyah piksel değerleri
..	..	
64	**6**	
65	**4**	
66	**2**	
67	0	Ara boşluk (siyah piksel içermeyen satırlar)
68	0	
69	0	
70	**6**	
71	**8**	Uzantı satıra (geçersiz satıra) ilişkin siyah piksel değerleri
72	**8**	
73	**7**	
74	0	
75	0	

Algoritma tarafından oluşturulan **r** matrisinin ilk ve son durumu :

1	66
70	73
85	112
…	…

⇨

1	**73**
85	112
…	…
…	…

SB işlemlerinde karşılan diğer bir zorluk da 'kef' گ harfi gibi, üst uzantısı bir önceki satırın alt sınırını ihlal eden harflerdir. Bu durumda ard arda gelen iki satır tek bir satır gibi algılanabilmekte ve SB işlemi hatalı sonuç vermektedir.

8.2.3 Satır Parçalama (SP)

Bir önceki bölümde anlatılan satır belirleme (SB) algoritmasıyla resim formatında taranmış belgedeki satırlar belirlenip bu satırlara ilişkin bilgiler (başlangıç, bitiş piksel satır numaraları) bir matriste saklanırken, satır parçalama (SP) algoritması ile de bu bilgileri kullanıp, her satır için benzer işlemler (satır içindeki kelime ve/veya harf gruplarını belirleme) yapılmış ve bu harf grupları resim dosyaları olarak kaydedilmiştir.

Algoritmanın girdileri :

I(m,n) : *mxn* boyutlarındaki resim tabanlı bir dokümana ilişkin ikili matris.
r : Dokümandaki satırların başlangıç ve bitiş piksel satır numaralarını içeren matris.

th : Satırda belirlenen harf grupları arasındaki boşluğun, en büyük boşluğa oranını belirleyen alt eşik değeri.

Algoritmanın yerel değişkenleri:

hist(n) : Satırların piksel sütunlarına ilişkin siyah piksel sayılarının tutulduğu dizi.
ch(k,2) : Satırda belirlenen **k** adet harf grubunun başlangıç ve bitiş piksel sütun numaralarının tutulduğu matris.

Algoritmanın çıktısı :

Dokümanın satırlarında belirlenen her kelime (veya harf grubu), kayıpsız bir resim dosyası (bmp) olarak saklanır.

r matrisini tarayarak, dokümandaki her bir satır için;

1. Satırın her piksel sütununa ilişkin siyah piksel sayılarını hesapla ve ***hist*** dizisinde sakla.

2. ***hist*** dizisinde başı ve sonu *0* ile belirlenmiş olan bloklarının başlangıç ve bitiş indislerini ***ch*** matrisinde sakla (bu bloklar siyah piksellere sahip piksel sütunları olup, satırdaki harf gruplarını temsil etmektedir).

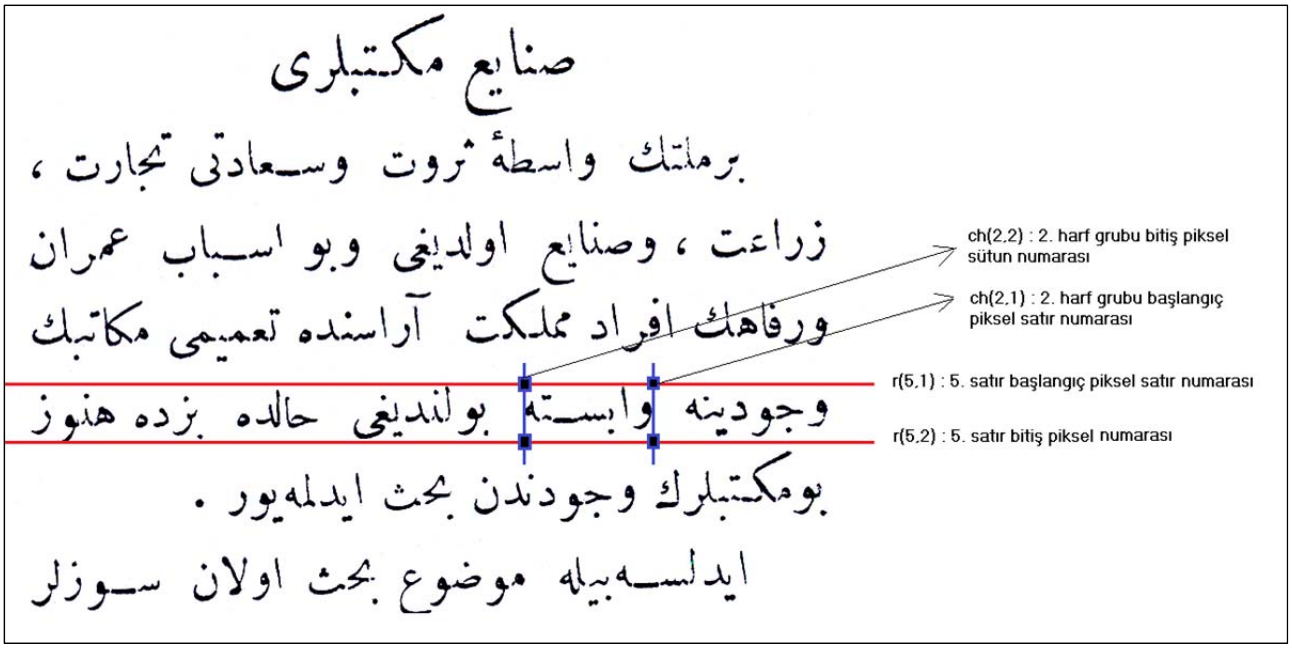
3. 2. adımında oluşturulan ***ch*** matrisini taramak suretiyle harf grupları arası boşlukların en büyük değerini hesapla (**ch(i,2)- ch(i+1,1)** değeri, satırın ***i***. ve ***i+1.*** harf grupları arasındaki boşluk uzunluğunun sütun piksel sayısı cinsinden değeridir).

4. ***ch*** matrisindeki her harf grubu arası boşluk için;

a. Harf grupları arası boşluk, en büyük boşluk uzunluğunun ***th*** katından daha büyük ise bu boşluğu geçerli kabul et.

b. Harf grupları arası boşluk, en büyük boşluk uzunluğunun ***th*** katından daha küçük ise bu boşluğu geçersiz kabul edip, iki harf grubunu birleştir.

5. **I** ikili resim matrisi, **r** satır bilgileri matrisi ve **ch** harf grupları matrisindeki bilgileri kullanarak satıra ilişkin belirlenen her harf grubuna ait alt matrisi bir resim dosyası olarak sakla.



Şekil 8.3 Satır Parçalama İşlemi

Örnek olarak Şekil 8.3'deki dokümanda belirlenmiş olan kelime (harf grubu), dokümanın 5. satırının 2. kelimesidir. MATLAB üzerinde tüm dokümana ilişkin sayısal ikili matrise **I** dersek, kelimeye ilişkin alt matris,

I(r(5,1):r(5,2),ch(2,2):ch(2,1))

şeklinde ifade edilir. Bu alt matris MATLAB'in *imwrite* fonksiyonuna parametre olarak verilerek kelimenin\harf grubunun bir resim dosyası olarak saklanması sağlanmış olur.

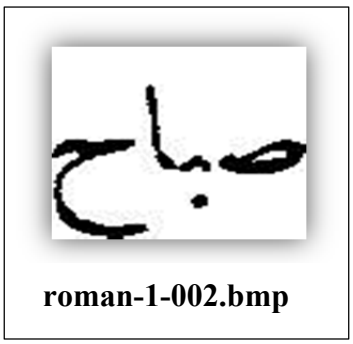
SP işlemiyle tek bir resimden oluşan her bir Osmanlıca belge, içerdiği harf grupları sayısınca küçük resim dosyalarına dönüştürülmüş olmaktadır. Satır parçalamadaki başarı oranı, sistemin doğru çalışması yada hata oranının az olmasında çok önemlidir. Bir sonraki adımda elde edilmiş olan harf gruplarının isimlendirilmesi (etiketleme) yapılacaktır.

8.2.4 Alan Etiketleme (Harf Gruplarının İsimlendirilmesi)

SP işlemiyle elde edilen harf gruplarının resimlerine; ait olduğu sınıf, belge numarası ve harf grubu numarasını içeren dosya isimleri verilmiştir. Harf gruplarının resimlerine verilen dosya isimleri;

sınıf_ismi-belge_no-sıra_no.bmp

formatındadır. (Örneğin; roman-1-002.bmp dosyası, roman sınıfına ait 1 numaralı belgeden elde edilmiş 2. harf grubuna ilişkin resim dosyasıdır Bkz.Şekil 8.4).



roman-1-002.bmp

Şekil 8.4 Örnek Harf Grubu Resmi ve Dosya İsmi

8.2.5 Benzerlik Matrisi

Çalışmamızın bir sonraki aşaması, harf gruplarının görüntülerini birbirleriyle karşılaştırarak, resimler arası benzerlik puanlarını hesaplamak ve benzer resimleri yani farklı yerlerde geçen aynı kelimeleri kümelemektir.

Osmanlıca'nın yapısal özelliğinden dolayı, yazı içerisinde karakterler kalınlık olarak değişebilmekte, aynı karakterin değişik konumlarda, değişik kalınlıkları olabilmektedir ya da bir karakter farklı formlarda yazılabilmektedir. Bu sebepten farklı boyutlardaki iki resmin, aynı kelimeyi içerse dahi sayısal görüntü matrisleri farklı olabilir ve bilgisayar bunları farklı resimler olarak algılayacağından yanlış etiketleme yapılmış olur. Bu da Dİ' nin, Osmanlıca için çözmesi gereken önemli problemlerindendir. O halde her bir resmin sayısal görüntüsünü temsil eden boyuttan bağımsız, özellik vektörlerine ihtiyaç vardır. Bu yüzden resimlerin özellik vektörleri tespit edilerek ikili kombinasyonlar halinde karşılaştırılır ve elde edilen benzerlik/farklılık puanları 'Benzerlik Matrisi' ne alınır.

Çalışmamızın harf gruplarından özellik çıkarma aşamasında Bölüm 5.1 de anlatılan yöntem [1] kullanılmış ve her bir resim iki farklı vektör (Dikey Tarama Vektörü, Yatay Tarama Vektörü) ile tanımlanmıştır.

8.2.5.1 YTV ve DTV Algoritmaları

- **Yatay Tarama Vektörü (YTV) :** Herhangi bir ikili resmin yatay tarama vektörü; resmi oluşturan piksel satırlarındaki toplam çizgi sayılarını içeren vektördür. Aynı satırda yer alan kesintisiz siyah piksel bloklarının her biri bir çizgi olarak nitelendirilir.

- **Dikey Tarama Vektörü (DTV) :** Herhangi bir ikili resmin dikey tarama vektörü; resmi oluşturan piksel sütunlarındaki toplam çizgi sayılarını içeren vektördür. YTV tanımındaki ile benzer şekilde; aynı sütunda yer alan kesintisiz siyah piksel bloklarının her biri bir çizgi olarak nitelendirmektedir.

8.2.5.2 Benzerlik Matrisinin Oluşturulması

Her bir resme ait özellik vektörleri (DTV - YTV) elde edildikten sonra resimler arası benzerlik puanlarının çıkarılması işlemine geçilmiştir. 'n' adet resimden oluşan veri setinde, ikişerli gruplar halinde tüm resimler birbirleriyle karşılaştırılarak resimler arası benzerlik/farklılık puanları (hata puanları - $\boldsymbol{P_{ij},\ 1 \leq i,j \leq n}$) hesaplanmış ve sonuçlar bir matris haline getirilmiştir. Benzerlik matrisi oluşturulurken sırasıyla aşağıdaki işlemler gerçekleştirilmiştir.

1. İlk resmin DTV si ile ikinci resmin DTV si arasındaki hata puanına $\boldsymbol{PD_{12}^0}$ diyelim. DTV vektörlerinin vektör elemanları karşılaştırılarak, farklı değerlere sahip elemanlar için hata puanına bir eklenerek $\boldsymbol{PD_{12}^0}$ hesaplandı.

2. İlk resmin YTV si ile ikinci resmin YTV si arasındaki hata puanı da $\boldsymbol{PY_{12}^0}$ olsun. $\boldsymbol{PY_{12}^0}$ de $\boldsymbol{PH_{12}^0}$ ile benzer şekilde hesaplandı.

3. İlk iki resmin DTV ve YTV leri +1 ve -1 yönde kaydırılarak, yukarıda anlatıldığı şekilde sırasıyla $\boldsymbol{PD_{12}^1}$, $\boldsymbol{PD_{12}^2}$, $\boldsymbol{PY_{12}^1}$, ve $\boldsymbol{PY_{12}^2}$ hesaplandı.

4. DTV ve YTV ler için hesaplanan üçer farklı hata puanından minimum olanlar alınarak toplandı. Ek olarak iki resmin DTV leri arasındaki uzunluk farkı ($\boldsymbol{D_{fark}}$) ve YTV leri arasındaki uzunluk farkı ($\boldsymbol{Y_{fark}}$) da eklenerek genel hata puanı ($\boldsymbol{P_{12}}$) hesaplandı.

$$\boldsymbol{P_{12} = \min(PD_{12}^m) + \min(PY_{12}^m) + D_{fark} + Y_{fark} \quad , \quad 0 \leq m \leq 2}$$

5. Aynı işlem ilk resimle tüm resimler arasında tekrarlanarak $\boldsymbol{P_{12}, P_{13}, P_{14} \ldots, P_{1n}}$ bulunur. Bu değerler benzerlik matrisimizin ilk satırını oluştururlar.

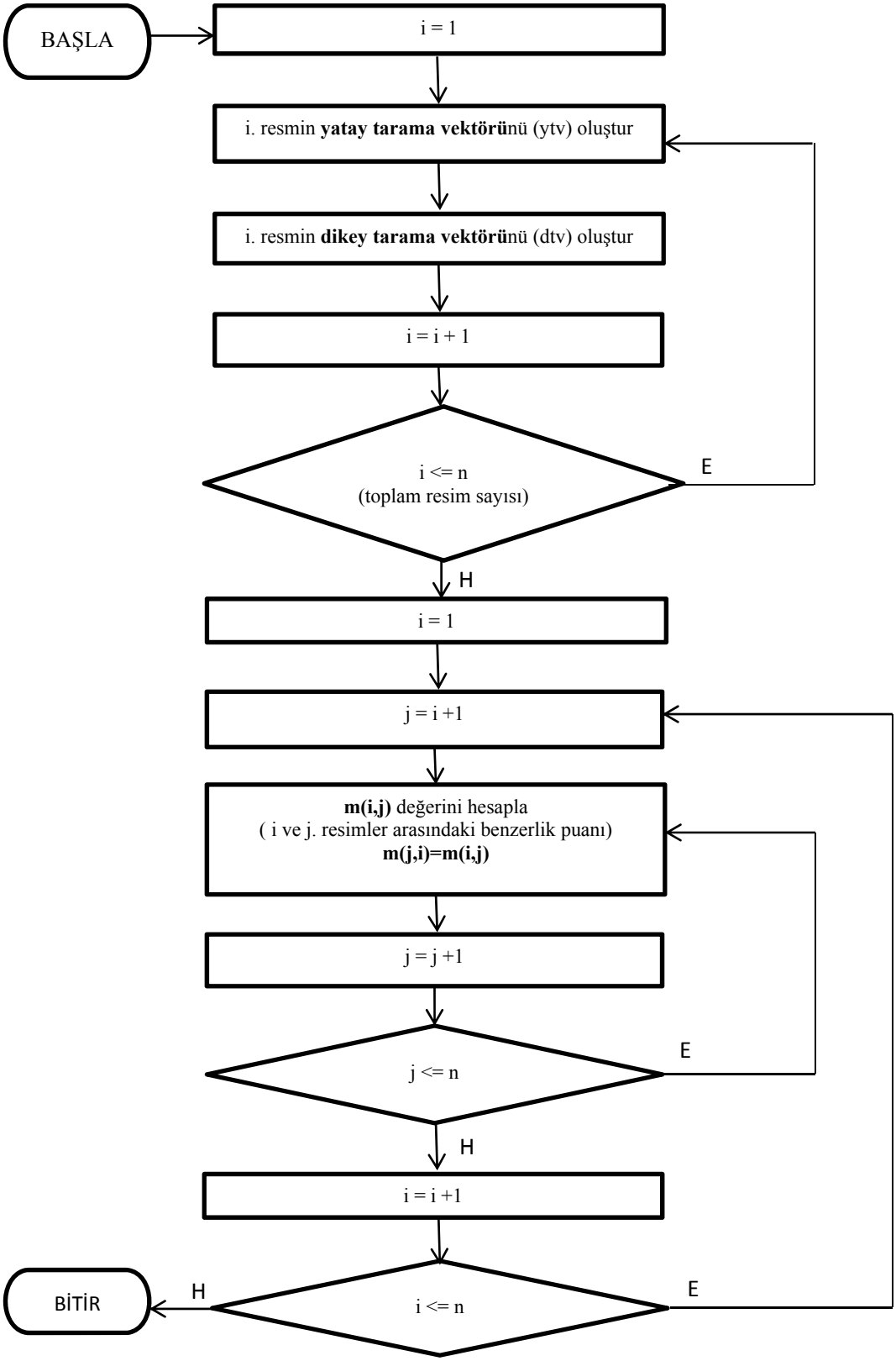
6. Matrisimizin simetrik bir kare matris ve asal köşegen elemanlarının 0 olduğu açıktır. $\boldsymbol{P_{jj} = 0}$ ve $\boldsymbol{P_{ij} = P_{ji}}$

7. Benzer şekilde $\boldsymbol{P_{i(i+1)}, P_{i(i+2)}, P_{i(i+3)}, \ldots, P_{in}}$ $(\boldsymbol{1 \leq i \leq n})$ puanları da hesaplanarak bulunan tüm değerler benzerlik matrisine ('BM') yerleştirildi (Bkz.Tablo 8.1).

Tablo 8.1 Benzerlik Matrisi

$$BM = \begin{bmatrix} 0 & P_{12} & P_{13} & P_{14} & \ldots & P_{1n} \\ P_{21} & 0 & P_{23} & P_{24} & \ldots & P_{2n} \\ P_{31} & P_{32} & 0 & P_{34} & \ldots & P_{3n} \\ P_{41} & P_{42} & P_{43} & 0 & \ldots & P_{4n} \\ . & . & . & . & . & . \\ P_{n1} & P_{n2} & . & . & P_{n(n-1)} & 0 \end{bmatrix}$$

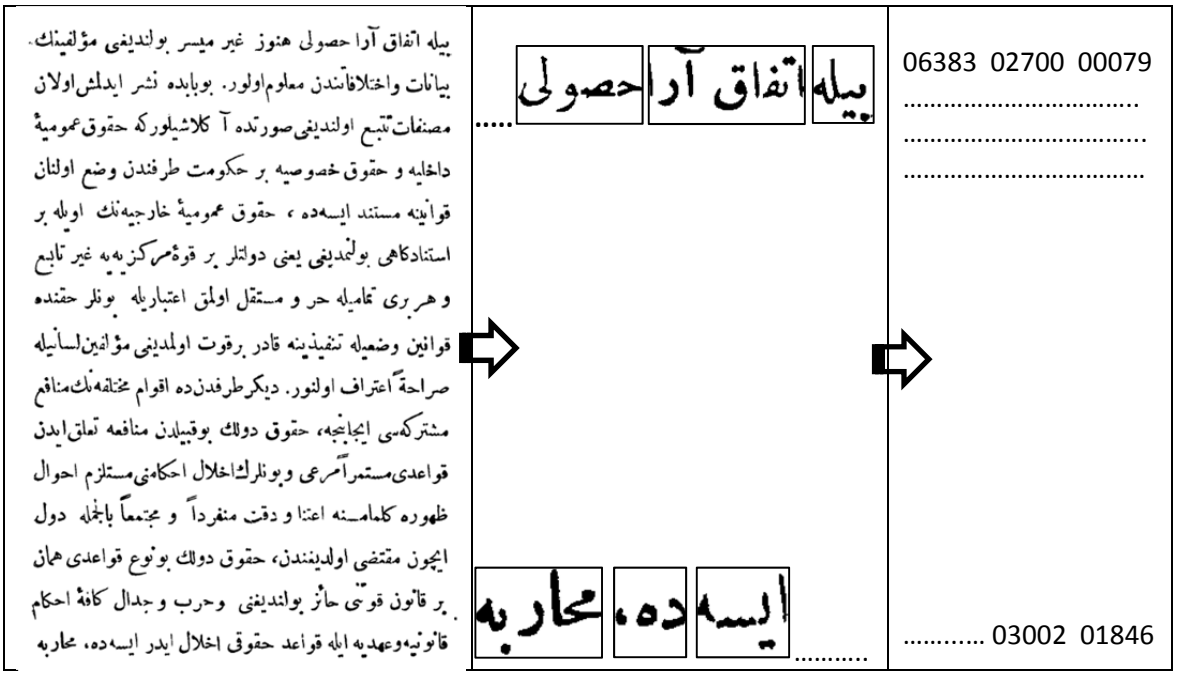
Benzerlik matrisi oluşturma algoritmasının akış diyagramı Şekil 8.5'de gösterilmiştir.



Şekil 8.5 Benzerlik Matrisi Şematik Algoritma Diyagramı

8.2.6 Kümeleme (Harf Gruplarının Kümelenmesi)

Çalışmamızda kümeleme için MATLAB paket programının hazır '*linkage (average)*' ve '*cluster*' fonksiyonları kullanılmıştır. Bu aşamada veri setinin tamamı bir havuzda toplanarak, birbirine benzer ya da aynı olan harf grupları kümelenmiş ve aynı kümedeki elemanların ortak birer küme numarası alması sağlanmıştır. Kümeleme işleminden önce yaklaşık 24000 resimden oluşan farklı küme sayısı, kümeleme sonrasında 7911 adede düşmüştür. Böylece veri setimizdeki aynı (veya benzer) harf gruplarının aynı sembolle ifade edildiği bir sayısal bilgi havuzu elde edilmiştir. Sonuç olarak roman, sosyoloji ve tarih sınıflarına ait belgeler kümeleme işlemiyle, resim formatından sayısal küme numaralarını içeren metin formatına dönüştürülmüştür (Bkz.Şekil 8.6).



Şekil 8.6 Osmanlıca Belgenin Harf Gruplarına Ayrıştırılması ve Küme Numaralarından Oluşan Metin Belgesine Dönüştürülmesi

8.2.7 Doküman Sınıflandırma Aşaması

Başlangıçta resim formatında bulunan Osmanlıca belgeler, yukarıda anlatıldığı şekilde harf grupları/kelimeler yerine sayısal küme numaralarını içeren metin belgelerine dönüştürüldükten sonra doküman sınıflandırma aşamasına geçilmiştir.

Bizim tezimize göre; daha önceki çalışmalarda metin formatındaki belgelerde uygulanan kelime frekans ya da n-gram istatistikleri ile sınıflandırma yöntemleri, resimlerin küme numaralarının yan yana yazılmasıyla elde edilen metin görünümlü belgelerde de çalışabilir. Metin formatındaki bir dokümanda aynı ASCII koduna sahip iki karakterin aynı harfi simgelemesi, resim formatlı bir dokümanda aynı küme numarasına sahip iki resmin aynı (veya çok benzer) harf gruplarını simgelemesi ile özdeştir.

Bu düşünceden hareketle, veri setinin tamamında harf gruplarının sayısal küme numaraları, ASCII kodları yerine kullanılarak harf gruplarının frekansları çıkarılmış ve dokümanların elde edilen özellik vektörleri ile sınıflandırma yapılmıştır.

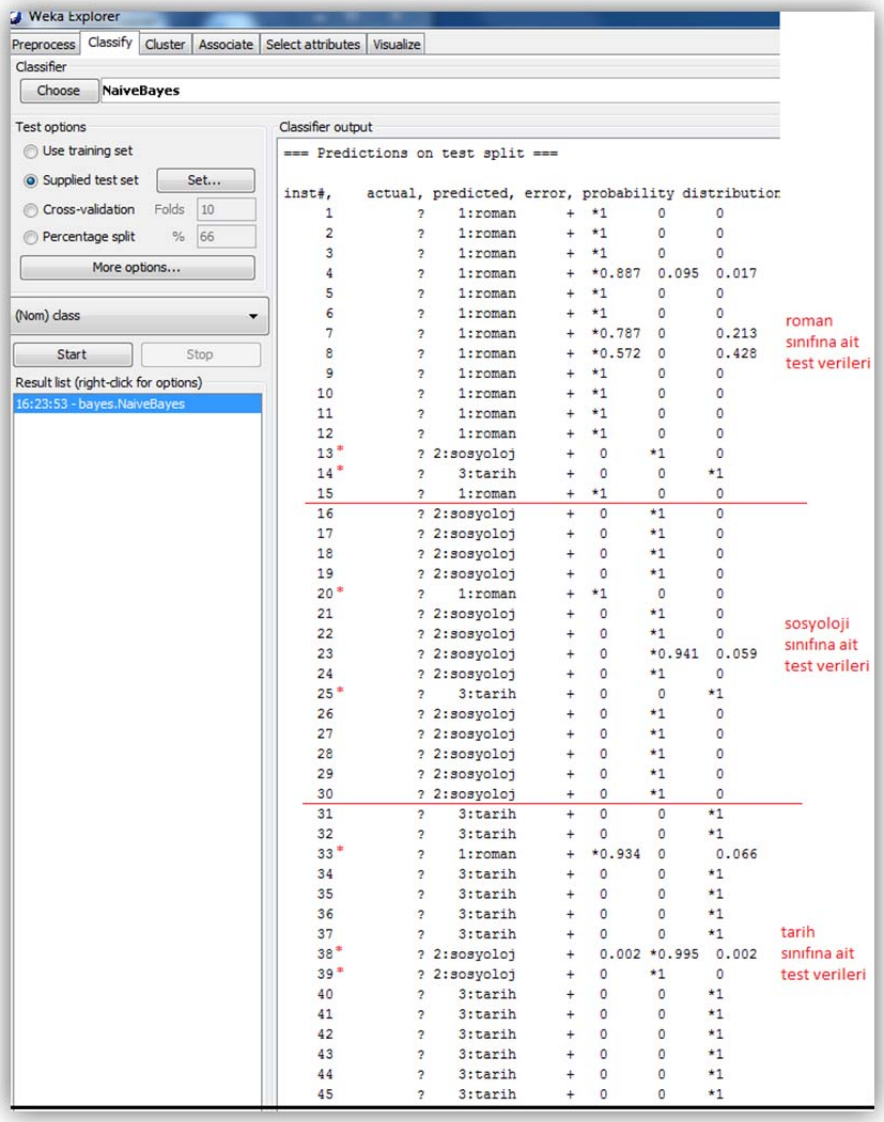
Veri setini oluşturan roman, sosyoloji ve tarih sınıflarına ait dokümanların 35'er tanesi eğitim verisi, kalan 15'er tanesi de test verisi olarak ayrılmıştır. Son olarak eldeki özellik vektörleri 'arff' formatına dönüştürülmüş ve WEKA programıyla sınıflandırma işlemi tamamlanmıştır. WEKA da temel sınıflandırma yöntemi olarak Naive Bayes ile aynı zamanda'Cross – Validation (CV)' ve 'Percentage Split' seçenekleriyle uygulamalar yapılmıştır. Ayrıca 'Attribute Selection' filtresi seçilerek ve seçilmeden ayrı ayrı sınıflandırmalar gerçekleştirilmiştir. Tüm çalıştırmaların karşılaştırmalı sonuçları Bölüm 9. da ayrıntılı olarak verilmiştir.

WEKA yapılan sınıflandırmaların haricinde, farklı bir sınıflandırma da şöyle yapılmıştır: Her dokümanın kelime frekanslarını içeren özellik vektörleri öklid uzaklık kullanılarak roman, sosyoloji ve tarih sınıflarının özellik vektörleriyle karşılaştırılmıştır. Test verisini oluşturan ve önceden sınıfı bilinmeyen dokümanlar bahsi geçen üç sınıftan kendisine en yakın olan sınıfa atanmıştır.

9. SONUÇ

9.1 Test Sonuçları

Weka programında yapılan sınıflandırmalarda sınıflandırıcı yöntem olarak 'Naive Bayes' seçilmiştir. İlk çalıştırmada roman, sosyoloji ve tarih sınıflarından 50'şer adet belgenin 35'er tanesi eğitim seti olarak girilerek sistem eğitildi ve bahsi geçen üç sınıfa ait 15'er adet belge 'supplied test set' seçeneği ile sisteme tanıtıldı, ardından sınıflandırma işlemi gerçekleştirildi. Toplam 45 belgeden oluşan test setinde 2'şer tanesi roman ve sosyoloji sınıflarından, 3 tanesi ise tarih sınıfından olmak üzere toplam 7 hata ile sınıflandırma tamamlanmıştır. Başarı oranı **%84,4** dür.



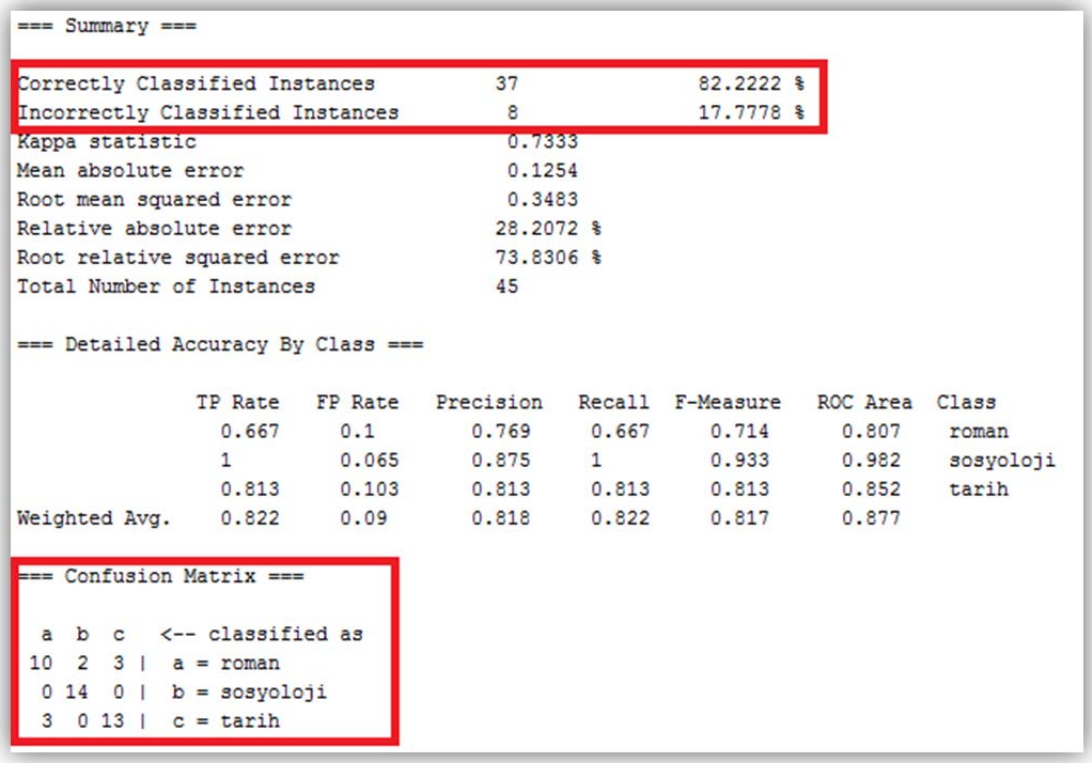
Şekil 9.1 'Suplied Test Set' (35+15) ile Sınıflandırma Sonuçları

İkinci çalıştırmada 10'lu 'Cross-validation (*CV(10)*)' seçeneği tercih edildi. Türkçe de 'Çapraz Doğrulama' olarak isimlendirilen bu yöntemde veri kümesi rastgele 'n' adet eşit alt gruba ayrılarak birinci grup test, diğerleri eğitim verisi olarak alınır ve sınıflandırma yapılır. İkinci adımda bir sonraki alt grup test, geriye kalan diğerleri eğitim verisi olarak alınır ve yeni bir sınıflandırma yapılır. Bu işlemler n tane alt grup için aynı şekilde tekrarlanır. Sistemin genel başarısı bu sonuçların ortalaması alınarak hesaplanır. Literatürde sıkça kullanılan 'n' değeri 10' dur. Çalışmamızda üç sınıftan 50 şer, toplamda 150 adet belge CV yöntemiyle sınıflandırıldı ve **%84** başarı oranı elde edildi. Elde edilen sonuçların ekran görüntüsü ve karşılaştırma matrisi Şekil 9. 2 de gösterilmiştir.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         126               84      %
Incorrectly Classified Instances        24               16      %
Kappa statistic                          0.76
Mean absolute error                      0.1045
Root mean squared error                  0.3165
Relative absolute error                 23.5039 %
Root relative squared error             67.1421 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.72      0.06       0.857     0.72      0.783      0.877    roman
                 0.94      0.09       0.839     0.94      0.887      0.957    sosyoloji
                 0.86      0.09       0.827     0.86      0.843      0.909    tarih
Weighted Avg.    0.84      0.08       0.841     0.84      0.838      0.914

=== Confusion Matrix ===

  a  b  c   <-- classified as
 36  7  7 |  a = roman
  1 47  2 |  b = sosyoloji
  5  2 43 |  c = tarih
```

Şekil 9.2 CV (10) ile Sınıflandırma Sonuçları

Üçüncü çalıştırmada veri setinin tamamı kullanıldı ve bu veri setinin Weka tarafından belirli bir oranda eğitim ve test verisi olarak ayrıştırılmasına olanak sağlayan 'Percentage split' seçeneği işaretlendi. Bizim uygulamamızda verinin %70 eğitim seti, %30 test seti olarak ayrıştırılması için gerekli parametre sisteme girildi ve sınıflandırma gerçekleştirildi. Test verisinin sınıf bilgileri, önceden sisteme verilmiş olduğu için sistem sınıflandırmayı yapar ve elindeki sonuçlara göre doğruluk oranını belirler. Bu çalıştırmada doğru sınıflandırılan belge oranı **%82.2** olarak bulundu. Elde edilen sonuçların ekran görüntüsü ve karşılaştırma matrisi Şekil 9.3'de gösterilmiştir.

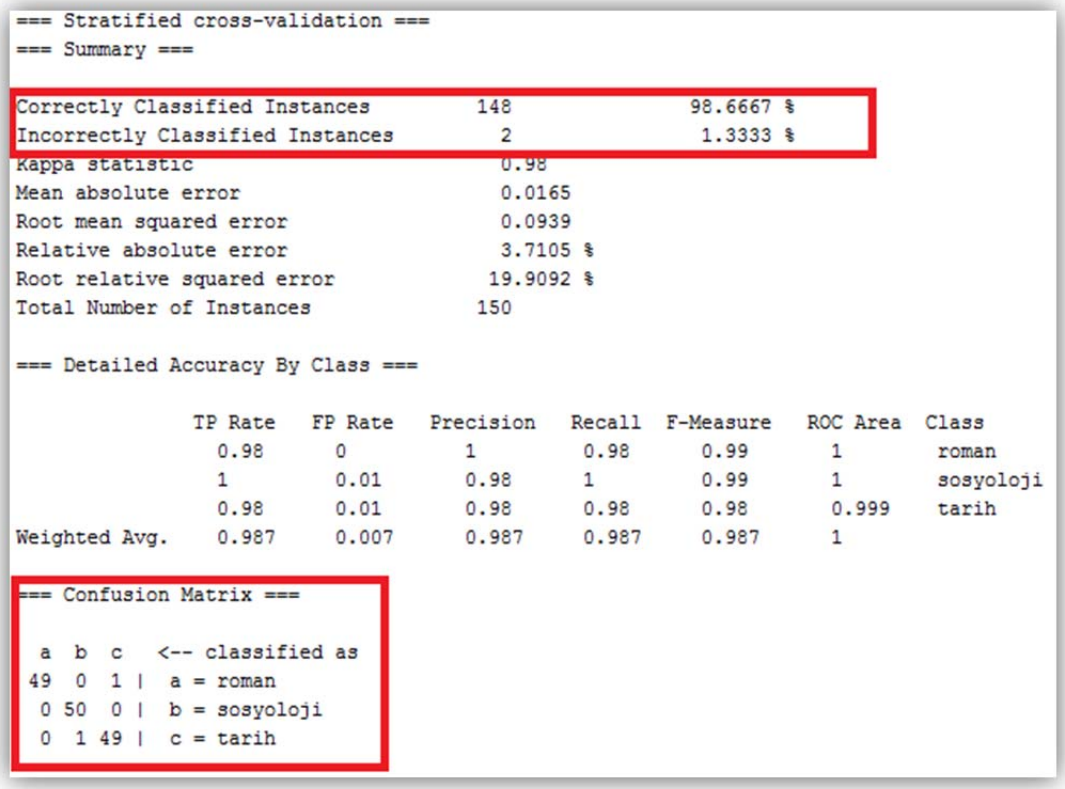


```
=== Summary ===

Correctly Classified Instances          37               82.2222 %
Incorrectly Classified Instances         8               17.7778 %
Kappa statistic                          0.7333
Mean absolute error                      0.1254
Root mean squared error                  0.3483
Relative absolute error                 28.2072 %
Root relative squared error             73.8306 %
Total Number of Instances               45

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.667    0.1       0.769     0.667     0.714      0.807    roman
                 1        0.065     0.875     1         0.933      0.982    sosyoloji
                 0.813    0.103     0.813     0.813     0.813      0.852    tarih
Weighted Avg.    0.822    0.09      0.818     0.822     0.817      0.877

=== Confusion Matrix ===

  a  b  c   <-- classified as
 10  2  3 |  a = roman
  0 14  0 |  b = sosyoloji
  3  0 13 |  c = tarih
```

Şekil 9.3 'Percentage Split' ile Sınıflandırma Sonuçları

Bundan sonraki çalıştırmalarda, ikinci ve üçüncü çalıştırmalardaki uygulamalar 'Attribute selection' düğmesi tıklanarak tekrarlandı. 'Attribute selection' özellik seçimi anlamına gelen bir filtredir. Diğerlerine göre daha az belirleyici olan etkisiz özellikleri tespit eder ve bu özellikleri eleyerek, daha az sayıda ama etkili olan özelliklerle sınıflandırma yapar. Böylelikle daha doğru sınıflandırma sağlar.

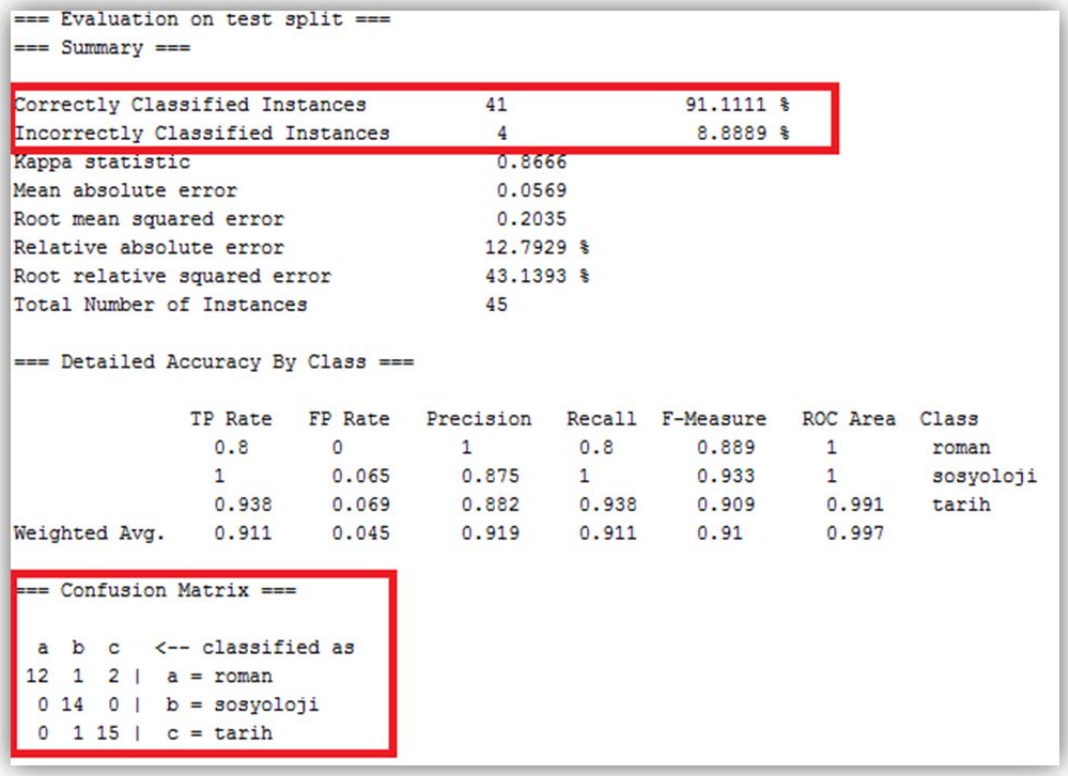
Dördüncü çalıştırmada CV (10), özellik seçim filtresiyle birlikte kullanıldı ve **%98.67** başarı oranına ulaşıldı. Sonuçların ekran görüntüsü ve karşılaştırma matrisi Şekil 9.4 de gösterilmiştir.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         148               98.6667 %
Incorrectly Classified Instances         2                1.3333 %
Kappa statistic                          0.98
Mean absolute error                      0.0165
Root mean squared error                  0.0939
Relative absolute error                  3.7105 %
Root relative squared error             19.9092 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.98      0          1         0.98      0.99       1        roman
                 1         0.01       0.98      1         0.99       1        sosyoloji
                 0.98      0.01       0.98      0.98      0.98       0.999    tarih
Weighted Avg.    0.987     0.007      0.987     0.987     0.987      1

=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  0  1 |  a = roman
  0 50  0 |  b = sosyoloji
  0  1 49 |  c = tarih
```

Şekil 9.4 CV (10) + 'Attribute Selection' ile Sınıflandırma Sonuçları

Beşinci çalıştırmada ‘Percentage split’ özellik seçim filtresi ile birlikte kullanıldı. Sınıflandırma başarı oranı **%91,1** olarak bulundu. Sonuçların ekran görüntüsü Şekil 9.5 gösterilmiştir.



```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          41               91.1111 %
Incorrectly Classified Instances         4                8.8889 %
Kappa statistic                          0.8666
Mean absolute error                      0.0569
Root mean squared error                  0.2035
Relative absolute error                 12.7929 %
Root relative squared error             43.1393 %
Total Number of Instances               45

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.8       0          1         0.8       0.889      1        roman
                 1         0.065      0.875     1         0.933      1        sosyoloji
                 0.938     0.069      0.882     0.938     0.909      0.991    tarih
Weighted Avg.    0.911     0.045      0.919     0.911     0.91       0.997

=== Confusion Matrix ===

  a  b  c   <-- classified as
 12  1  2 |  a = roman
  0 14  0 |  b = sosyoloji
  0  1 15 |  c = tarih
```

Şekil 9.5 ‘Percentage Split’ + ‘Attribute Selection’ ile Sınıflandırma Sonuçları

Weka ile yapılan testlerin karşılaştırmalı sonuç tablosu Tablo 9.1 de verilmiştir.

Tablo 9.1 Weka'da Yapılan Testlerin Karşılaştırmalı Sonuçları

Yöntem	Şekil No	Sınıf I : Roman		Sınıf II : Sosyoloji		Sınıf III: Tarih		Gen.Baş. Oranı
		+	-	+	-	+	-	
Suplied Test Set	9.1	13	2	13	2	12	3	%84,4
CV- (10)	9.2	36	14	47	3	43	7	%84
Percentage Split %70-%30	9.3	10	5	14	0	13	3	%82,2
CV (10) + Attribute Selection	9.4	49	1	50	0	49	1	% 98,67
Percentage Split + Attribute Selection	9.5	12	3	14	0	15	1	% 91,1

Yapılan sınıflandırma çalışmalarında Tablo 9.1 de görüleceği gibi en yüksek başarı oranı 'CV (10) + Attriute Selection' yöntemiyle elde edildi. Kullanılan yöntemler 'Attribute Selection' filtresiyle birlikte alındığında sınıflandırma başarı oranının arttığı gözlemlenmektedir. Sistem eğitildikten sonra girilen test verileri ile sınıflandırma yapan 'Suplied test set' seçeneğinde de modelimizin oldukça iyi bir başarı oranı ile çalıştığı gözlenmiştir.

WEKA da yapılan sınıflandırmaların haricinde, MATLAB ortamında gerçeklenen ve harf grubu frekans analizi ile sınıflandırma yapan yöntemimizde, eğitim setindeki her bir belgedeki harf gruplarının frekansı birer vektörde tutulmuş ve aynı sınıfa ait belgeler için ilgili vektörlerin ortalaması sınıfı temsil eden bir özellik vektörü olarak elde edilmiştir. Ardından test verilerinin özellik vektörleri sınıf özellik vektörleri ile öklid uzaklığı bazında karşılaştırılmış ve ilgili belgeler en yakın sınıflara atanmıştır. Bu yöntemde %69' luk başarı oranı yakalanmıştır.

9.2 Tartışma ve Öneriler

Bundan önceki çalışmalarda metin formatındaki belgelerde uygulanan doküman işleme ve doküman sınıflandırma yöntemleri bu tezde resim formatındaki Osmanlıca belgelere uygulanmıştır. Metin formatlı belgelerde, kelimeler bilgisayar tarafından ASCII kodlarıyla tanınmasına karşılık bu çalışmada harf gruplarının resimleri bilgisayara tanıtılmış ve her resim adeta bir karakter gibi algılatılarak, kelime frekans analizi yerine harf grubu frekans analizi ile doküman sınıflandırma yapılmıştır. Böylece savunduğumuz tezin Osmanlıca arşiv belgelerinde çalışabildiği gösterilmiştir. Elde edilen yüksek başarı oranları modelin doğru ve geliştirilebilir olduğunu kanıtlamaktadır.

Ülkemizin sahip olduğu zengin Osmanlı arşivlerinin önemli bir kısmının elektronik ortama aktarılmış olmasına rağmen tasnifin hala elle yapıldığı düşünüldüğünde, sunduğumuz çalışmanın ve ileride geliştirilerek kullanıma uygun hale getirilmesinin önemi ortaya çıkmaktadır. Özellikle el yazısı belgelerde sınıflandırma yapabilen modellerin geliştirilmesi de çok önemli bir ihtiyaçtır.

Kurduğumuz modelde, anlamsal (semantik) analiz yapılmadan belgenin görüntüsü üzerinden sınıflandırma yapıldığı için ortaya konan bu model Osmanlıca haricinde Arapça, Çince, İbranice gibi farklı dillerde de çalışabilir.

Model hazırlandıktan sonra uygulamanın çalıştırılması, İntel Core i.7, 3.4 GHz işlemcili ve 8,0 GB ana bellek kapasitesine sahip bir makinayla 5 saatte tamamlanmıştır.

Doküman işleme aşamasında, belge sayısı ile parça sayısı (harf gruplarının resim sayısı) doğru orantılı olarak artar ancak kümeleme safhasında parça sayısı ile benzerlik matrisinin boyutu karesel olarak artar. Bizim çalışmamızda 150 adet belgede yaklaşık 24.000 parça elde edildiği için benzerlik matrisi 24.000 x 24.000 boyutlarında olup 576.000.000 eleman içermektedir. Daha büyük hacimlerdeki veri setlerinde muhtemel performans problemlerini ortadan kaldırmak için paralel programlama teknikleri uygulanabilir.

Uygulamada karşılaştığımız ve ileriki çalışmalarda geliştirilebileceğine inandığımız diğer noktalar şunlardır:

- Model birkaç modülden oluşur (SB, SP , Kümeleme, BM ve Sınıflandırma gibi). Her modülün kendi içinde ihmal edilebilecek küçük hatalar, zincirleme birbirlerine eklenerek sistemin bütününde çarpımsal oranda büyür ve genel bir hataya sebep olabilir.

- Doküman sınıflandırmada öncelikle 'Kelime 2-gram' tekniği seçilmiş fakat uygulamada istenilen başarı oranına ulaşılamadığı görülmüştür. Bunun nedeni olarak; kümeleme aşamasında bazı özel durumlarda, aynı iki harf grubunun aynı küme numarasını almamasının sonuca etkisinin büyük olduğu tespit edilmiştir. Ayrıca kelime-gram tekniğinin başarısı veri hacminin büyüklüğü ile doğru orantılıdır. Veri setinin hacmi arttıkça bir kelime grubunun birlikte geçme sıklığı (frekans) artacak böylece dokümanın bilgisayar tarafından doğru tanınma olasılığı daha fazla olacaktır.

- Coğrafya, Anatomi, Matematik vb. gibi şekil ve resim içeren dokümanlarda bu şekillerin ayrıştırılabilmesi için özel çözümler üretilmelidir.

- Bu tezde matbu dokümanlar üzerinde çalışılmıştır. El yazısı ile yazılmış dokümanlarda, yazı karakterleri yazan kişiye göre değiştiğinden ve satırların çoğunlukla düz bir hat üzerinde olmamasından dolayı SB ve SP daha zor çalışacaktır. Bunun için dile özgü kurallar işletilerek parçalamanın daha sağlıklı yapılması sağlanabilir.

10. KAYNAKLAR

[1]. Chew L. T., Member, IEEE Computer Society,Weihua H., Zhaohui Yu, and Yi Xu ‘‘Imaged Document Text Retrieval Without OCR’’ IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, No. 6, 2002.

[2]. Khreisat, L., ‘‘Arabic Text Classification Using N-Gram Frequency StatisticsA Comparative Study’’ Conference on Data Mining 2006, Dept. of Computer Science, Math and Physics Fairleigh Dickinson University.

[3]. Yalnız İsmet Z., İsmail Şengör Altıngövde,Uğur Güdükbay,Özgür Ulusoy, ‘‘Integrated segmentation and recognition of connected Ottoman script’’, Optical Engineering 48(11), 117205 (November 2009), Bilkent University Department of Computer Engineering.

[4]. Tan, Chew L., Huang, W., Sung, Sam,Y., YU, Z., Xu,Y., ‘‘Text Retrieval from Document Images Based on Word Shape Analysis’’, Applied Intelligence 18, 257–270, 2003, Kluwer Academic Publishers.

[5]. Bespalov,D., Bai, B., Qi, Y., ‘‘Sentiment Classification Based on Supervised Latent n-gram Analysis’’ NEC Labs America, CS Dept, Drexel University.

[6]. Khreisat L., ‘‘A machine learning approach for Arabic text classification using *N*-gram frequency statistics’’ , Journal of Informetrics 3 (2009) 72–77.

[7]. Takçı, H., Güngör, T., ‘‘A high performance centroid-based classification approach for language identification’’ , Pattern Recognition Letters 33 (2012) 2077–2084

[8]. Polat S., Başbakanlık Müsteşar Yardımcısı, ‘‘Önsöz’’, Başbakanlık Osmanlı Arşivi Rehberi, T.C. BAŞBAKANLIK DEVLET ARŞİVLERİ GENEL MÜDÜRLÜĞÜ Osmanlı Arsivi Daire Başkanlığı Yayın Nu: 42, İkinci Baskı İstanbul-2000, Erişim Tarihi 01.08.2013, <http://www.devletarsivleri.gov.tr/Forms/pgArticle.aspx?Id=0f2a5cfb-c614-4f67-9c7f-db18cb167ed1>.

[9]. Amasyalı, M. F., Balcı,S., Varlı, E., N., Mete, E., ‘‘ Türkçe Metinlerin Sınıflandırılmasında Metin Temsil Yöntemlerinin Performans Karşılaştırılması’’, Elektrik Elektronik Fakültesi, Bilgisayar Mühendisliği Bölümü Yıldız Teknik Üniversitesi.

[10]. Gharib, T. F., Habib,M. B., Fayed, Z. T., ‘‘Arabic Text Classification Using Support Vector Machines’’, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.

[11]. Suen, C.Y., Bergler,S., Nobile, N., Waked,B., Nadal,C.P., Bloch, A., ‘‘Categorizing Document Images Into Script And Language Classes’’ Proc. Int'l Conf. on Advances in Pattern Recognition, ICAPR'98, 1998.

[12]. El-Halees,A.M., ‘‘Arabic Text Classification Using Maximum Entropy’’ The Islamic University Journal (Series of Natural Studies andEngineering)Vol. 15, No.1, pp 157-167, 2007, ISSN 1726-6807,

[13]. <http://bbytezarsivi.hacettepe.edu.tr/jspui/handle/2062/132> Erişim Tarihi: 28.07.2013

[14]. Vega, F.S., Tello, E.V., Gomez, M. M., ‘‘Determining and characterizing the reused text for plagiarism detection’’ Elsevier 2012, www.elsevier.com/locate/eswa Erişim Tarihi : 28.08.2013

[15]. Doğan S.,‘‘Türkçe Dokümanlar İçin N-Gram Tabanlı Sınflandırma:Yazar Tür ve Cinsiyet’’,Yüksek Lisans Tezi, Yıldız Teknik Üniv., Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bölümü, İstanbul 2006.

[16]. Sanan, M., Rammal, M., Zreik, K., ‘‘ Arabic document classification using N-gram’’IEEE, Paris 8 University, Paris, France, Lebanese University, Beirut, Lebanon.

[17].<http://www.arsivder.org.tr/alt2-kategori.asp?id=150&sayfa=32&grup=Osmanlıca &isim =Osmanlıca Nedir?> Erişim Tarihi: 29.07.2013

[18].<http://tr.wikipedia.org/wiki/Naive_Bayes sınıflandırıcı> Erişim Tarihi: 30.07.2013

[19]. Nabiyev, V.V., “Yapay Zeka”, Seçkin Yayıncılık, Ankara, 506 -525 (2012).

[20]. Mahmoud, R., Majed, S., ‘‘Improving Arabic Information Retrieval Systemusing n-gram method’’ , WSEAS Transactions on Computers, Volume 10 Issue 4, Pages 125-133, April 2011.

[21]. Diri B., Doğan S., ‘‘Türkçe Dokümanlar İçin N-gram Tabanlı Yeni Bir Sınıflandırma(Ng-ind): Yazar, Tür ve Cinsiyet’’ Türkiye Bilişim Vakfı Bilgisayar Mühendisliği Dergisi, Sayı 3, Haziran 2010.

[22]. Keikha,M., Khonsari, A., Oroumchian, F.,‘‘Rich document representation and classification: An analysis’’ ELSEVIER , Knowledge-Based Systems 22, 67-71,2009

[23]. Srihari N. S., Ball R.G., Srinivasan H., ‘‘Versatile Search of Scanned Arabic Handwriting’’ Center of Excellence for Document Analysis and Recognition (CEDAR),University at Buffalo, State University of New York,Amherst, New York 14228

[24]. Marwan A. H.,Omer M.S.L., ‘‘Stemming Algorithm To Classify Arabic Documents’’ , Symposium on Progress in Information & Communication Technology 2009.

[25]. Amasyalı, M. F., Diri, B., “Automatic Turkish Text Categorization in Terms of Author, Genre and Gender”, 11th International Conference on Applications of Natural Language to Information Systems-NLDB2006, LNCS Volume 3999, 2006.

[26]. Çiltik, A. ve Güngör, T., "Time-Efficient Spam E-mail Filtering Using N-gram Models", Pattern Recognition.

[27]. Aras, P., ''Bilgisayar Destekli El Yazısı Karakterlerini Tanıma Sistemi Tasarımı'' Yüksek Lisans Tezi, İstanbul Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Haziran 2006.

[28]. Günay Atbaş, A. C., ''Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma'', Yüksek Lisans Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Ana Bilim Dalı, Ankara 2008.

[29]. Eroğlu, Y., ''Osmanlıca El Yazısı Harfleri Çevrim İçi Tanıma'', Yüksek Lisans Tezi, Gazi Üniversitesi, Bilişim Enstitüsü, Elektronik-Bilgisayar Eğitimi Bölümü, Ankara, Temmuz 2007.

[30]. Özhan, D., ''Osmanlıca Karakterlerin Yapay Sinir Ağları İle Tanınması'', Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik Eğitimi Bölümü, Ankara, Aralık 2005.

[31]. <http://tr.wikipedia.org/wiki/Ar%C5%9Fiv#.5Bhttp:2F.2Fwww.devletarsivleri.gov.tr.2F_Ba.C5.9Fbakanl.C4.B1k_Osmanl.C4.B1Ar.C5.9Fivi.5D> Erişim Tarihi: 28.07.2013

[32]. Caballero, F.A., Lopez, M. T., Castillo, J. C., ''Display text segmentation after learning best-fitted OCR binarization parameters'', ELSEVIER, Expert Systems with Applications 39, 4032–4043, 2009.

[33]. Peng F., Keselj V., Cerconey N., Thomasy C., (2003), "N-Gram-Based Author Profiles For Authorship Attribution", Faculty of Computing Science, Dalhousie University, Canada.

[34]. Diri B., Amasyalı, M.F., (2003), "Automatic Author Detection for Turkish Text", 13th International Conference on Neural Information Processing, Turkey.

[35]. Alpaydın, E., Akın, L., Aratma, S., Yagcı, M., " Yapay Sinir Agları İle Görüntü Tanıma", TÜBİTAK Proje, EEEAG-41, Ankara, 8-15. (1994).

[36]. <http://tr.wikipedia.org/wiki> Erişim tarihi 22.08.2013

[37]. <http://www.bilgisayarkavramlari.com > Erişim tarihi: 22.0 8.2013

[38]. < http://www.yildiz.edu.tr/~bayram/sgi/saygi.htm> Erişim tarihi : 28.07.2013

[39]. <www.ist.yildiz.edu.tr/dersler/dersnotu/Kum-Analiz.doc> Erişim tarihi : 22.0 8.2013

[40]. Anderberg M.R. 1973. ''Cluster Analysis for applications''. Academic Press, New York. Page 553–555.

[41]. George H., 1995, "Estimating Continuous Distributions in Bayesian Classifiers", Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338-345. Morgan Kaufmann, San Mateo.

[42]. <http://www.tbmm.gov.tr/develop/owa/e_yayin.liste_q?ptip=EHT> Erişim Tarihi: 28.07.2013

[43]. Zeng J., Wu C., Wang W., ''Multi-grain hierarchical topic extraction algorithm for text mining'' , ELSEVIER, Expert Systems with Applications 37, 2010, 3202–3208.

[44]. Akşehirli Ö:Y:, Ankaralı H.,Aydın D.,Saraçlı Ö., ''Tıbbi Tahminde Alternatif Bir Yaklaşım:Destek Vektör Makineleri'', Turkiye Klinikleri Journal of Biostatistics 2013 - Volume 5 Issue 1,19-28.

[45] <http://www.kemik.yildiz.edu.tr/data/File/egiticili_agirliklandirma.pdf> Erişim Tarihi: 19.09.201