BAŞKENT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DOKÜMAN KATEGORİZASYONU VE İMZA BÖLGE ANALİZİ

İLKHAN CÜCELOĞLU

YÜKSEK LİSANS TEZİ

2014

DOKÜMAN KATEGORİZASYONU VE İMZA BÖLGE ANALİZİ

DOCUMENT CATEGORIZATION AND SIGNATURE REGION ANALYSIS

İLKHAN CÜCELOĞLU

Başkent Üniversitesi

Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin

BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü

YÜKSEK LİSANS TEZİ

Olarak hazırlanmıştır.

2014

“Doküman Kategorizasyonu ve İmza Bölge Analizi” başlıklı bu çalışma, jürimiz tarafından, 13 / 08 / 2014 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI’nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan : Prof. Dr. Adnan Yazıcı

Üye (Danışman) : Doç. Dr. Hasan Oğul

Üye : Yrd. Doç. Dr. Mustafa Sert

ONAY

/ 08 / 2014

Prof. Dr. Emin AKATA

Fen Bilimleri Enstitüsü Müdürü

TEŞEKKÜR

Sayın Doç. Dr. Hasan Oğul'a her zaman yardımcı, yol gösterici olduğu ve bana büyük zamanlar kazandırdığı için..

Bilim, Sanayi ve Teknoloji Bakanlığı'na ve DAS A.Ş.'ne San-tez kapsamında destek oldukları için.. (Proje No : 01522.STZ.2012-2)

Gülşen Avcı, Bahadır Şükrü Yılmaz ve Murat Yüksel'e test amaçlı kullanılacak dokümanların sağlanması ile ilgili yardımları için ..

ÖZ
DOKÜMAN KATEGORİZASYONU VE İMZA BÖLGE ANALİZİ

İlkhan CÜCELOĞLU

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Bu tezde, taranmış doküman görüntülerinin otomatik analizi üzerine çalışmalar yapılmıştır. Bu amaçla doküman analizinde iki alt problem ele alınmıştır; dokümanların otomatik kategorizasyonu ve doküman üzerinde imza tespiti. Doküman tabanlı resimlerin kategorizasyonu birçok uygulama için önemli bir araçtır. Bu çalışma bankacılık uygulamalarında sık kullanılan dokümanları kategorize eden bir altyapıyı tanıtmaktadır. Altyapı, dokümandan oluşturulan metin bilgisi ve doküman resim özniteliklerini kullanmaktadır. Özniteliklerin çıkartılması ve seçilmesi ile ilgili teknik uygulanmış ve Türkçe metinler için özelleştirilmiştir. Dokümanın resim özniteliklerini kullanarak yapılan kategorizasyon ise, işlem maliyeti yüksek olan optik karakter tanıma işlemine gereksinim duymadığından daha hızlı sonuç veren bir alternatif sunmaktadır.

Dokümanlarda elle atılan imzanın bulunduğu bölgenin otomatik olarak belirlenmesi bankacılık, sigorta ve kamu sektöründeki iş süreçlerinde katma değer üretebilecek bir özelliktir. Çalışma, herhangi bir tip sigorta dokümanından imzanın çıkarılmasını sağlayan bir altyapıyı tanıtmaktadır. Geliştirilen altyapı, bölütlere ayrılmış resmin temsil eden resim öznitelikleri ile sınıflandırılması işlemine dayanmaktadır. Bölütleme, iki etaplı bağlı bileşenlerinin etiketlenmesi ile gerçekleştirilmektedir. Bölütler, farklı öznitelik temsil yöntemleri ile vektöre çevrilip, destek vektör makineleri ile sınıflandırılarak imza içeren ve içermeyen olarak ayrıştırılmaktadır. Gerçek sigorta dokümanlarından oluşan veri kümesi üzerinde yapılan deneyler, geliştirilen altyapının yüksek doğruluk değerlerine ulaşabildiğini ve gerçek hayattaki uygulamalarla birlikte çalışabileceğini göstermektedir.

ANAHTAR SÖZCÜKLER : Doküman analiz ve tanıma, resim işleme, bilgisayarla görme, örüntü tanıma

Danışman : Doç.Dr. Hasan OĞUL, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

ABSTRACT

DOCUMENT CATEGORIZATION AND SIGNATURE REGION ANALYSIS

İlkhan CÜCELOĞLU

Başkent University Institue of Science and Engineering

Computer Engineering Department

This thesis contains studies related to automated analysis of document images. Two sub-problems in document analysis are considered for his purpose; automated categorization of documents and handwritten signature detection on documents. Classifying document images is an essential tool for many applications. This work presents a framework for categorizing documents which are frequently used in bank applications. The framework is based on the extracted text information and document image features. A feature extraction and selection technique is applied customized for Turkish texts. Categorization based on document image features is an alternative giving results in a faster way, because it works without the optical character recognition process, which is a computational intensive task.

Automated localization of a handwritten signature in a scanned document is a promising facility for many banking and insurance related business activities. This work also describes here a discriminative framework to extract signature from a insurance service application document of any type. The framework is based on the classification of segmented image regions using a set of representative features. The segmentation is done using a two-phase connected component labeling approach. The combined effects of several feature representation schemes in distinguishing signature and non-signature segments is evaluated over a Support Vector Machine classifier. The experiments on a real insurance data set have shown that the developed framework can achieve a reasonably good accuracy to be used in real life applications.

KEYWORDS : Document recognition and analysis, image processing, computer vision, pattern recognition.

Advisor : Assoc. Prof. Dr. Hasan OĞUL, Başkent University, Department of Computer Engineering.

İÇİNDEKİLER LİSTESİ

ŞEKİLLER LİSTESİ

ÇİZELGELER LİSTESİ

SİMGELER VE KISALTMALAR LİSTESİ

DPI	Dot Per Inch
HOG	Histogram of Gradients
LBP	Local Binary Patterns
LTP	Local Ternary Patterns
MICR	Magnetic Intelligent Character Recognition
MNB	Multinomial Naive Bayes
OCR	Optical Character Recognition
SIFT	Scale Invariant Feature Transform
SURF	Speeded Up Robust Features
SVM	Support Vector Machine

1. GİRİŞ

1.1 Motivasyon

Kurumların günümüzün hızlı değişen dünyasında işlerini başarılı bir şekilde sürdürebilmeleri için doğru bir şekilde yönetilmeleri ve oluşan tüm verileri dikkate alabilmeleri gerekmektedir. Bu durum sadece bilgi sistemlerinde bulunan sayısal verilerin değil, diğer ortamlarda bulunan yapısal olmayan verilerin de iş süreçleri ile entegre edilmesi sonucunu ortaya çıkarmaktadır.

Günümüzde kurumlara ait verilerin yüzde yirmisinin yapısal formatta, yüzde sekseninin ise yapısal olmayan formatta olduğu değerlendirilmektedir [1]. Yapısal olan veriler veritabanları ile saklanmakta ve bilgi sistemleri ile yönetilmektedir. Yapısal olmayan veriler ise doküman, e-posta, faks vb. verilerden oluşmakta olup, bu verileri elektronik ortamda yönetmek için doküman, içerik yönetim sistemleri ve arşiv sistemleri kullanılmaktadır.

Dokümanlarda bulunan verilere elektronik ortamda erişilebilmesi, öncelikle fiziksel dokümanın elektronik ortamda oluşturulmasını gerektirmektedir. Tarayıcılar ile dokümanların taranarak elektronik ortama aktarılması bu aşamada kullanılan temel yöntemdir. Faks ile gönderilen dokümanlarda, kullanılan yazılımlar ile elektronik ortamda oluşturulabilmektedir. Doküman tarayıcılara ek olarak, günümüzde mobil cihazlardan fotoğraf çekme yöntemi ile oluşturulan resimler de işlenmeye başlanmıştır.

Dokümanların elektronik ortama aktarılması kurumlara birçok fayda sağlamaktadır. Bunlar arasında dokümanlara daha kolay erişim, iş süreçlerindeki süre ve maliyetin azaltılması, fiziksel dokümanların depolanması için daha az alana ihtiyaç duyulması, çalışan verimliliğinin artması, doküman güvenliğinin artması, doküman gönderme işleminin daha esnek, hızlı ve ucuz olması sayılabilir [2].

Tez kapsamında, elektronik ortamda bulunan resim tabanlı dokümanlar üzerinde çalışılmaktadır. Resim üzerinden gerekli verilerin alınabilmesi için doküman kalitesi ve özellikleri önem taşımaktadır. Tarayıcıların sürücüleri içerisinde bulunan bazı özellikler ile taranan dokümanlar üzerinde görüntü iyileştirme özellikleri yapılabilmektedir. Bu işlemler, elektronik ortamda oluşan dokümanın görüntü

kalitesini ve bilgi çıkarma işlemlerindeki başarı oranını artırmaktadır. Faks yolu ile gelen dokümanlarda ise, doküman kalitesi tarayıcılar ile oluşturulan dokümanlara göre daha kötü olabilmektedir. Yazının silik olabilmesi, satırların eğri bir şekilde oluşması, kenarlarda siyah bölgeler oluşması, sayfa alınırken oluşan kaymalardan ötürü yazı kalitesinin bozulması dokümanlar elektronik ortama aktarılırken oluşabilecek sorunlar arasındadır. Elektronik ortama aktarım sırasında kullanılan çözünürlük, doküman kalitesi için ayrı bir belirleyici özelliktir. Hem okunabilirliğin iyi olması, hem de optik karakter tanıma (OCR) işlemleri için üç yüz DPI çözünürlük değeri olan uygulamada yaygın olarak kullanılmaktadır.

Dokümanlar elektronik ortamda oluşturulduktan sonra, dizinleme işlemi yapılır. İndeksleme, dokümanla ilgili üstverilerin oluşturulması ve dokümanla ilişkili bir şekilde kaydedilmesi işlemidir [3]. Dokümanla ilgili üstveriler kurumun ilgili dokümana erişmek için kullanacağı verilere göre değişkenlik göstermektedir. Dokümanın tarihi, konusu, ilgili kurum gibi bilgiler örnek olarak gösterilebilir. İndeksleme sürecinde verilerin otomatik olarak alınması tercih edilmekte, verinin bu şekilde alınamaması durumunda kullanıcılar tarafından oluşturulması gerekmektedir.

Kullanıcı tarafından elle girilen bu bilgilerden bir tanesi de doküman grubunun belirlenmesidir. Doküman grubu bir dokümanın hangi sınıfa ait olduğunu belirtmekle birlikte, hangi üstverilerin girilmesi gerektiğini, nasıl bir depolama yöntemi ile saklanması gerektiğini, doküman hayat döngüsünü, dokümanın kullanılabileceği iş akışlarını da belirlediği için doküman yönetim sistemlerinde kullanılan en önemli kavramlardan biridir. Doküman grubuna göre kullanılacak üstverilerin değişmesine örnek olarak, faturalar doküman grubunda tarih, tutar, firma adı gibi üstveriler kullanılırken, nüfus cüzdanı doküman grubunda kimlik numarası, ad, soyad gibi üstverilerin kullanılması örnek olarak gösterilebilir. Bazı doküman grupları kritik bilgi içerip değiştirilmez olduklarından değiştirilmez medyaya kaydedilirken, bazı dokümanlar ise sabit disklere kaydedilmektedir. Doküman hayat döngüsünün bir süreci olan arşivleme evresinde, bazı doküman grupları beş yıl, bazıları ise on beş hatta yüz yıla kadar saklanabilmektedir. Son olarak, bir dokümana ait doküman grubunun belirlenmesi, dokümanın girmesi gereken iş akışını belirlenip otomatik olarak başlatılmasını sağlayabilmektedir.

Örnek olarak günümüzde yoğun olarak kullanılmakta olan iş süreç yönetim sistemlerinde, doküman yönetim sistemine bir izin talep dokümanı kaydedildiği zaman, otomatik olarak izin iş sürecinin başlatılması gibi kurallar konabilmektedir.

Doküman gruplarının yapısı hiyerarşik olarak oluşturulabileceği gibi, tek bir seviye içerecek şekilde oluşturulabilir. Kimlik dokümanları altında, nüfus cüzdanı ehliyet, pasaport gibi alt gruplar olabileceği gibi, bu gruplar tek seviye halinde doküman yönetim sistemlerinde oluşturulabilirler.

Doküman yönetim sistemlerine doküman girdisinin oluşturulmasını sağlayan doküman yakalama (capture) sistemlerinde ise, doküman grupları dokümanların resim özelliklerini ve üzerlerinde yapılacak resim bazlı işlemleri belirlemek açısından önem kazanmaktadır. Dokümanın hangi çözünürlükte taranacağı, renk derinliğinin ne olması gerektiği, üzerinde yapılması gereken resim iyileştirme işlemleri, sıkıştırma formatı gibi çeşitli bilgiler yine doküman grubu kavramı ile yönetilmektedir. Doküman grubunun yazılımlar tarafından otomatik olarak belirlenmesini sağlayan doküman kategorizasyon işlemi, doküman yakalama işlemlerinde de süreçleri iyileştirerek verimliliği arttırabilecek özellikler taşımaktadır.

Birçok doküman tabanlı süreçte büyük bir öneme sahip olan doküman grubunun doğru olarak belirlenmesi kritik bir önem taşımaktadır. Tez konusu olan doküman sınıfının belirlenmesi günümüzde kullanıcılar tarafından gerçekleştirilmekte olup, bu durum iş süreçlerini aşağıdaki sebeplerden daha verimsiz hale getirmektedir :

- Sınıflandırma işleminin iş süreçlerinde kullanılan her doküman için yapılması gerektiğinden zaman kaybı oluşmaktadır. İşlem süresinin artması, iş süreçlerindeki daha verimsiz hale getirmekte, müşteri ve çalışan memnuniyetini azaltmaktadır.
- Yüksek bilgi giriş maliyeti yüzünden bazı dokümanlar dosya bazında dizinlenmekte, doküman bazında dizinlenmemektedir. Dosyada birden çok doküman grubu olması durumunda, veriye erişim zor ve maliyetli bir hale gelebilmektedir. Örnek olarak, bankada hesap açılış sürecinde kullanılan kimlik dokümanı ayrıca sınıflandırılmaz ve başvuru dosyası içerisine bu şekilde kaydedilir ise, ilgili başvurunun kimlik fotokopisine ulaşmak

istendiğinde tüm dosyanın ilgili sisteme indirilmesi ve sayfalar arasından kimlik fotokopisinin bulunması gerekmektedir.

- Bir işlemin gerçekleştirilmesi için gereken zorunlu doküman grupları var ise, bu doküman yerine yanlışlıkla başka bir doküman kaydedilerek işlem devam ettirilebilmektedir. Bilgisayar yazılımları asıl doküman tipine uygun bir doküman ile işleme devam edildiğini kontrol etmediğinden, tarafından yapılan ikinci bir kontrol bulunmamaktadır.
- Kullanıcıların ilgili süreçlerde kullanılan doküman sınıflarını öğrenmesi gerekmektedir. Çalışanların değişmesi durumunda yeni eğitim verilmesi gerekmekte ve maliyet oluşmaktadır. Yeni gelen personelin deneyimli personel seviyesinde iş üretimine geçmesi zaman almaktadır. Bu durum, doküman sınıflandırma işleminde doğruluğun dönemsel olarak azalabilmesini beraberinde getirmektedir.

Geniş bilgi sistemleri ile iş süreçlerini entegre eden bankacılık sektöründe, iş süreçlerinde kullanılan dokümanların sınıflandırılması giderek önemini arttıran önemli bir iş haline gelmiştir. Doküman gruplarının yazılım tarafından doküman bazında belirlenebilmesi önem kazandığı gibi, bir doküman kümesi içerisinde belirli bir doküman sınıfına ait dokümanın bulunup bulunmadığı da belirlenmek istenmektedir. Örnek olarak toplu olarak taranan bir doküman kümesinde nüfus cüzdanı fotokopisi bulunup bulunmadığını öğrenmek gerekebilmektedir. Tez kapsamında bankacılık sektöründe iş süreçlerinde yoğun olarak kullanılan on dokuz doküman sınıfı seçilmiş ve bu doküman sınıflarının sınıflandırılması için bir altyapı sunulmuştur.

Tez kapsamında üzerinde çalışılan bir diğer konu, dokümanlar üzerinde elle atılmış bir imza olup olmadığının belirlenmesi işidir. Ele alınan dokümanlar tarayıcılar ile oluşturulmuş ve genellikle siyah/beyaz renkler kullanılarak bir bit renk derinliğinde kaydedilmiştir. Dokümanların arşivde daha az yer kaplaması, depolama maliyetlerinin azaltılması siyah beyaz doküman kullanımı için sebeplerden bir tanesidir. Buna ek olarak, bu dokümanların boyutlarının daha küçük olması, gerektiği zaman ağ üzerinden aktarımı daha hızlı bir hale getirmektedir. Bu kapsamda, birçok doküman üzerinden daha çok bilgi alınabilmesi için renkli olarak taransa bile, bilgiler alındıktan sonra siyah beyaz

renklere çevrilerek arşivlenmektedir. Bu duruma örnek olarak taranan banka çekleri örnek olarak gösterilebilir. Çekler, renkli olarak taranmaktadır. Çek üzerinde banka kodu, şube kodu, hesap numarası gibi bilgiler özel bir font kullanılarak çek basımı sırasında oluşturulmaktadır. Bankalar bu işlem için genellikle E13B fontunu kullanmaktadır [44]. Bu fontun otomatik olarak bilgisayar tarafından tanınması işlemi donanım seviyesinde veya yazılım seviyesinde yapılmakta, bu işlem için taranan çek resminin renkli olarak oluşturulması gerekmektedir. Bilgi alındıktan sonra, çek resminin merkeze iletilmesinden önce, çekler gri tona çevrilmektedir. Böylece iletim, depolama ve daha sonraki görüntüleme işlemleri için performans ve yer avantajı elde edilmektedir.

Gerek elektronik arşivlerdeki oluşturulan yasal dokümanların çoğunlukla siyah beyaz olması, gerekse birçok tarama işleminin siyah/beyaz renklerle yapılması imza bölge analizi konusunda renk bilgisi kullanımı yerine başka yaklaşımlar kullanılarak imzanın dokümanda yer tespitinin yapılmasını gerekli hale getirmiştir. Bu durum faks yolu ile gelen dokümanlara ait iş süreçlerinde de geliştirilen altyapının kullanılabilmesini mümkün kılmıştır.

Siyah beyaz dokümanlar üzerinde imza bölge analizinin yapılabilir duruma gelmesi kurumlara birçok fayda sağlamaktadır. Öncelikle kurumlar bazı iş süreçlerini yasal anlamda geçerli dokümanlar ile başlatmak zorundadır. Örnek olarak bir eve elektrik bağlatmak için bir dilekçe veya başvuru dokümanı oluşturulmalıdır. Bu dokümanın yasal anlamda geçerli olabilmesi için, dokümanda elle oluşturulmuş bir imza bulunması gerekmektedir. Kullanıcı hatası, dikkatsizlik vb. sebeplerden iş süreçleri dokümanda imza olmadan başlatılabilmektedir. Dokümanda imzanın bulunmaması sonradan oluşabilecek hukuki sorunlarda kurum için zararla sonuçlanabilmektedir. Bir önceki elektrik bağlanması örneğinde, kurum yasal olarak geçersiz bir başvuru ile elektrik bağlatmış ve elektrik hizmetini belirli bir süre vermiş olabilir. İlgili abone hizmet ücretini ödemediğinde imzası alınmadığı için, kurum verdiği hizmetlerle ilgili zarara uğramış olacaktır.

İmza bölge analizinin arşiv dokümanları üzerinde gerçekleştirilmesi ve imzanın dokümandan ayrıştırılabilmesi, arşivler üzerinde imza doğrulama işlemi projelerini de gerçekleştirilebilir hale getirmektedir. İmza doğrulama işlemi, bir imzanın bir kişiye ait olup olmadığını belirlemektedir. Bu işlem için kişinin imzasının sistemde

bulunması gerekmektedir. Sistemdeki imza referans alınarak bir başka imzanın bu kişiye ait olup olmadığı belirlenmektedir. İşlemdeki doğruluğun (accuracy) yüksek olması için, sorgulanmak istenen imza ilgili dokümandaki diğer bilgilerle kesişmeyecek şekilde kesilmelidir. Bu işlem kullanıcılar tarafından gerçekleştirilmesi kullanılan güncel dokümanlarda maliyet oluşturmaktadır. Aynı zamanda, elektronik arşivlerde bulunan yüksek sayıda dokümanın her sayfası için bu işlemi yapmak çok maliyetli olduğundan, arşivlerdeki dokümanlar üzerinde imza doğrulma işlemi yapılamamaktadır. Önerilen alt yapı ile, bu projelerin de yapılabilir hale gelmesi beklenmektedir.

İmza bölge analizi uygulamaları bu faydalara ek olarak, çok sayfadan oluşan dokümanlarında, imzanın hangi sayfada olduğunu bularak birçok süreçte iyileştirme sağlamaktadır. Özellikle çağrı merkezleri, banka şubeleri gibi müşterilere ait atılan imzaların gözle kontrol edilmesinin gerektiği süreçlerde imzayı içeren dokümanın bütün sayfaları ile açılması, ağ trafiği arttırdığı gibi sayfaların gözle kontrolünü gerektirmektedir. Bu süreçteki maliyeti azaltmak için, güncel uygulamalarda imza bulunan sayfalara barkod eklenmiş, barkodun pozisyonuna göre imza atılan kutuların otomatik olarak kesildiği uygulamalar geliştirilmiştir. Ancak bu yöntemde de hatalar oluşabildiği gibi, buna ek olarak, arşivdeki dokümanlar için ilgili yöntem kullanılamamaktadır. İmza bulunan sayfanın yazılımlar tarafından otomatik olarak belirlenmesi bu açından da kritik bir uygulama haline gelmiştir.

Tez kapsamında geliştirilen altyapı, sigorta sektöründe kullanılan başvuru dokümanları üzerinde imza bulunan sayfanın belirlenmesi sağlamaktadır. Geliştirilen altyapıda verilen dokümandaki tüm sayfalar ayrıştırılmakta ve tüm sayfalarda aranmaktadır. Başvuru dokümanların basım tarihi, formatı vb. özellikler dikkate alınarak imzanın bulunabileceği sayfaları belirleyen sigorta sektörüne özel kuralların konulması durumunda, imza tüm sayfalarda değil, belirlenen alt kümede aranacağından başarı oranının bu yaklaşımda artabileceği değerlendirilmektedir. Ek olarak, imza bulunması durumunda altyapının oluşturduğu bir başka bilgi olan imzanın sayfa bulunduğu pozisyon verileri, imzanın dokümandan ayrıştırılması işlemi için kullanılabilecektir.

1.2 Tezin Katkıları

Doküman kategorizasyonu ve imza bölge analizi konularında tezin katkıları bulunmaktadır. Bununla birlikte, birbirinde ayrı gibi görünen bu iki konunun aslında birlikte çalıştığı zaman ek katma değer sağlayan bir altyapı ürettiği de görülmüştür.

İmza bölge analizinde, dokümanın sayfalarında imza olup olmadığı belirlenmektedir. Gereksinimlere genel olarak bakıldığında, dokümanda belirli bir imzanın bulunması ve varlığının kontrol edilmesi gerekmektedir. Örnek olarak bir başvuru dokümanında başvuruyu yapan kişinin imza atıp atmadığı kontrol edilmek istenmekte, başvuru ile ilgili diğer dokümanlarda bulunan imzalar ile ilgilenilmemektedir. Başvuru dosyasında ise, başvuru dokümanı, nüfus cüzdanı fotokopisi vb. çeşitli belgeler bulunmaktadır. Altyapı, nüfus cüzdanı fotokopisi ve başvuru sayfasında üzerinde bulunan imzayı bulabilmekte, imzanın kime ait olduğu veya anlamı ile ilgili bir çalışma yapılmamaktadır. Bu yüzden imza bölge analizi çalışması yapılmadan önce dosya içinde bulunan dokümanların kategorize edilmesi, gereksinimleri karşılamak açısından önem kazanmıştır.

Dokümanların resim tabanlı sınıflandırılması yıllardır doküman analizi ve tanıma üzerinde çalışan araştırmacıların büyük ilgisini çekmiş, bu konuda birçok çalışma yapılmıştır [4,5,7,8,9,10,11,12,13]. Konu ile ilgili literatürde yapılmış olan çalışmaların özeti giriş bölümünde bahsedilmektedir. Bununla birlikte, tez kapsamında yapılan çalışma ile bankacılık sektöründe yoğun olarak kullanılan Türkçe dokümanların kategorizasyonu ile ilgili bir çalışma bulunmamaktadır. Bankacılık süreçlerinde kullanılan gerçek dokümanlar ile bir küme oluşturulması, yöntemlerin bu küme üzerinde sonuçlarını görmek açısından önemlidir. Metin özellikleri ile doküman kategorizasyonu yönteminde, Türkçe desteği olan OCR motorları kullanımının doküman kategorizasyonuna etkisi incelenmiştir. Bütünce oluşturulurken kelimenin tamamının veya belirli bir kısmının alınmasının sonuçlara etkisi gözlenmiştir. Resim özellikleri ile doküman kategorizasyon yönteminde, LTP (Local Ternary Patterns) kullanılmıştır. Ayrıca bağlı bileşenlerin histogramına dayanan bir yaklaşımı oluşturulmuş, bu ve diğer yöntemlerin sonuçları destek vektör makineleri kullanılarak incelenmiştir. Global özellikler ile birlikte kullanıldığında, her iki yöntemin yakın bir performans gösterdiği görülmüştür.

Sonuçlar, tez kapsamında oluşturulan alt yapının iş hayatında kullanılan gerçek uygulamalara fayda sağlayabileceğini göstermektedir. Gerçek dokümanlardan oluşturulan veri kümesinde, metin tabanlı doküman kategorizasyonunda %89,42 doğruluk (accuracy) elde edilmiştir. Resim tabanlı doküman kategorizasyonunda ise, aynı veri kümesi üzerinde %69,87 doğruluk elde edilmiştir. Metin tabanlı doküman kategorizasyonunun uygulama alanı, doküman/içerik yönetim sistemlerine kaydedilmiş dokümanlardır. Bu uygulamada, altyapı sunucular üzerinde çalışacaktır. Resim tabanlı doküman kategorizasyonu ise, OCR işlem maliyetinden dolayı metin tabanlı yöntemin uygulanamayacağı durumlarda, daha düşük işlem maliyeti ile bir alternatif oluşturmaktadır. Özellikle tarama anında dokümanın kategorisinin belirlenmesi, doküman kalitesinden dolayı OCR başarısının düşük olduğu faks dokümanlarının kategorize edilmesi gibi örnekler bu yaklaşımın örnek uygulama alanlarıdır.

İmza analizi doküman analizi ve tanıma alanında araştırmacıların büyük dikkatini çekmiş olmasına rağmen, bu çalışmalar imzanın doğrulanması yönünde zenginleşmiş, imzanın dokümanda yerinin tespit edilmesi konusunda çok çalışma yapılmamıştır. Üzerinde çok çalışma yapılmamış olmakla birlikte, imza bölge analizi çeşitli zorluklar içermektedir. Bu duruma ek olarak, literatürde bulunan çalışmalarda daha az sayıda dokümandan oluşan veri kümeleri üzerinde denemeler yapılmıştır. Tez kapsamında yapılan çalışma ile, iş süreçlerinde kullanılan gerçek dokümanlardan oluşan bir veri kümesi oluşturulmuştur. Bu veri kümesinde 9943 sayfa bulunmakta olup, bu sayfaların tamamı üzerinde yapılan testler, yöntemlerin yüksek sayıda gerçek doküman üzerinde başarı oranlarını ortaya koyması açısından önem taşımaktadır.

İmza bölge analizi çeşitli öznitelik temsil yöntemleri ile gerçekleştirilmiş ve yöntemlerin karşılaştırmalı sonuçları raporlanmıştır. Bu aşamada SIFT (Scale Invariant Feature Transform), HOG (Histogram of Gradients), Gradyan yöntemi, global öznitelikler, LTP (Local Ternary Patterns) gibi yöntemler ve çeşitli kombinasyonları denenmiştir. 9943 sayfa üzerinde yapılan testlerde %71 doğruluk elde edilmesi, oluşturulan altyapının gerçek dokümanlar üzerinde iş uygulamalarından kullanılabileceğini göstermektedir.

İmza bölge analizi çalışmasında karşılaşılan problemli durumlar belirlenmiş, seçilen bazı örnekler raporlanmıştır. Böylece, doğruluğu arttırmak için ilerde yapılacak çalışmalar için geliştirilen yaklaşımın çözüm getirmediği durumlar netleştirilmeye çalışılmıştır.

Tezin geri kalan kısmı şu şekilde organize edilmiştir :

İkinci kısımda doküman kategorizasyonu ile ilgili çalışma yer almaktadır.

İlk bölümde, literatürde bu alanda yapılan çalışmalar özetlenmiştir.

İkinci bölüm, kullanılan yöntemi içermektedir. Doküman kategorizasyonu için kullanılan metin tabanlı öznitelikler, resim tabanlı öznitelikler ve iki her iki öznitelik temsilinin birlikte kullanıldığı hibrid yaklaşım açıklanmıştır. Bu bilgilerden sonra, sınıflandırma yöntemi ve yöntem ile ilgili açıklama bulunmaktadır.

Üçüncü ve son bölüm doküman kategorizasyonu ile ilgili sonuçları içermektedir. Metin öznitelikleri, resim öznitelikleri ve hibrid yaklaşım kullanılarak elde edilen sonuçlar ayrı ayrı listelenmiştir.

Üçüncü kısımda imza bölge analizi ile ilgili çalışma yer almaktadır.

İlk bölüm olan giriş bölümünde, imza analizi ile ilgili literatürde bulunan çalışmalar özetlenmiş, taranan dokümanlar imza bulma ile ilgili zorluklar belirtilmiştir.

İkinci bölüm kullanılan yöntemleri içermektedir. Ön işleme, bölütleme, kullanılan öznitelik çıkarım yöntemleri açıklanmıştır. Sınıflandırma ile ilgili bilgiler bu bölümde bulunmaktadır.

Son bölümünde sonuçlar bulunmaktadır. Sonuçlar iki ayrı veri kümesi üzerinde test edilmiş olup, her iki küme hakkında bilgi verilmektedir. Geliştirilen altyapının gerçek dokümanlar üzerindeki performansını gözlemleyebilmek için oluşturulan veri kümesi, küme bilgilerini ve test sonuçları içeren veritabanı yapısı açıklanmaktadır. Her iki veri kümesi üzerinde sonuçlar, sonuçlar ile ilgili analiz, oluşan problemler ile ilgili örnekler bölümün son kısmında bulunmaktadır.

Dördüncü kısımda çalışma ile ilgili genel sonuç değerlendirmesini ve gelecek çalışmalar ile ilgili analizi içermektedir.

2. DOKÜMAN KATEGORİZASYONU

2.1 Giriş

Resim tabanlı dokümanların sınıflandırılması yıllardır doküman analizi ve tanıma üzerinde çalışan araştırmacıların büyük ilgisini çekmiştir. Farklılıklarına rağmen, tüm metotlar üç aşamadan oluşan bir yapı kullanır: öznitelik çıkarımı, özniteliklerin temsili ve öğreticiyle öğrenme ile çıkarım. Metotlar yoğun olarak bu aşamalarda kullanılan tekniklere göre farklılık göstermektedirler.

Dokümanların sınıflandırılmasında bir başka ayırıcı özellik ise uygulama alanıdır; burada iş yazıları, faturalar, vergi/sigorta/banka formları, kitap veya dergi sayfaları sayılabilir. İlk çalışmalarda, faturaları ve vergi formlarını, yapısal özellikler ile bilgi tabanlı bir karar verme metodu kullanarak bir yaklaşım sunulmuştur [4]. İş yazılarını sınıflandırmak için metin özniteliklerini kullanan sinir ağları tabanlı bir yapı kullanılmıştır [5]. Dergi sayfalarının sınıflandırılması için doküman resim özniteliklerine saklı Markov modeli adapte edilmiştir [6]. Dokümanın fiziksel düzeni bir öznitelik temsili olarak sigorta formlarını sınıflandırmak için kural tabanlı bir karar sisteminde kullanılmıştır [7]. Dergi sayfalarının sınıflandırılması, fiziksel düzen ile sinir ağları [8] ve metin öznitelikleri ile Rocchio algoritması [9] şeklinde iki farklı çalışmada incelenmiştir. Fiziksel düzen bazı başka çalışmalarda en yakın k komşu yöntemi ile kitap sayfalarını [10] ve saklı Markov modeli ile faturaları sınıflandırmak için kullanılmıştır [11]. Banka dokümanları için bir deneme şablon eşleştirme tekniği ile yapılmıştır [12]. Resim öznitelikleri üstünde çok örnekli öğrenme (multiple instance learning) kullanan genel bir sınıflandırıcı sunulmuştur [13].

Doküman sınıflandırma işlemi tarayıcı ve kamera sistemlerinden elde edilen düşük kaliteli ve gürültülü içerikten oluşan resimlerde sorun yaşamaktadır. Bir başka zorluk ise, şablon seviyesinde benzerliği genellikle yüksek olan doküman tiplerinin çeşitliliğidir. Örnek olarak, elektrik, su ve doğal gaz harcamaları için ev faturaları şablon ve renk içeriği açısından çok benzemektedirler. Burada bahsedilen ikinci zorluk, resimleri temsil ederken metin içeriğini de kullanan, Türk bankalarında sık kullanılan bankacılık dokümanlarını sınıflandıran bir metot geliştirerek çözülmeye çalışılmıştır. Gerçek veri üzerinde yapılan deneyler altyapının farklı tipteki

bankacılık dokümanlarını ayrıştırma yeteneğini göstermektedir. Çalışma ayrıca farklı nitelik temsillerinin ve problem için sık kullanılan iki makine öğrenimi sınıflandırıcısının karşılaştırmalı analizini sunmaktadır.

2.2 Yöntem

2.2.1 Metin öznitelikleri

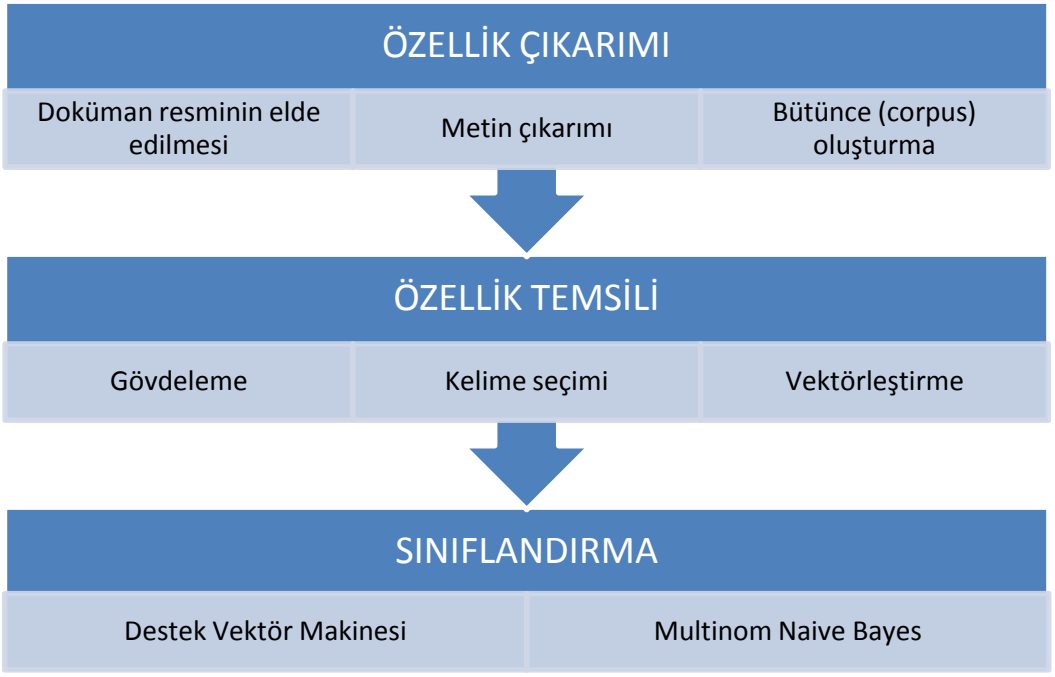
Metot

Resim tabanlı doküman sınıflandırma için kullanılan metin tabanlı yönteme ait genel altyapı Şekil 1 ‘de gösterilmiştir. Geliştirilen altyapı, öznitelik çıkarımı, öznitelik temsili ve sınıflandırma süreçlerinden oluşmaktadır.

Doküman resminin elde edilmesinden sonra Türkçe metin desteği olan bir optik karakter tanıma (OCR) motoru kullanılmıştır. OCR motoru çıktı olarak bir metin dosyası üretmektedir. Çıkarılan metin dosyasından sonra bir dizi işlem uygulanarak bütünce (corpus) oluşturulmuştur. Bütünce bilgilerinin veritabanına kaydedilmesi ile birlikte öznitelik çıkarımı adımı sona ermektedir.

Öznitelik temsili sürecinde ilk adım, gövdeleme işlemine göre bütüncede bulunan kelimelerin düzenlenmesidir. İşlemin uygulanması durumunda, kelimelerin ilk dört harfi seçilir. Sonraki adım, öznitelik vektöründe kullanılacak kelimelerin seçilmesidir. Bu işlemden sonra, çıkarılan metin, terim frekansına bağlı olarak bir öznitelik vektörüne çevrilmiştir.

Sınıflandırma için sık kullanılan iki öğreticiyle öğrenme tekniği ele alınmıştır : destek vektör makineleri ve Multinom Naive Bayes (MNB).



Şekil 1 : Resimden metin tabanlı doküman sınıflandırma için altyapı

Öznitelik çıkarımı ve temsili

Doküman resmi, metin içeriğini tanımlayan öznitelik kümesi ile temsil edilir. Bu noktada, doküman resminden metin içeriğini çıkartabilmek için optik karakter tanıma (OCR) motoru kullanılır. OCR işlemi için iki farklı motor kullanılarak sonuçlar karşılaştırılmıştır. Kullanılan iki motor Tesseract [42] ve Abby OCR [43] motorudur. Tesseract, açık kaynak kodlu bir motor olup, Google tarafından geliştirilmektedir. Abby OCR motoru ise, Abby firması tarafından geliştirilmiş, başarı oranı sektörde en yüksek olarak bilinen ticari bir OCR motorudur.

Tüm dokümanların çıkartılmış metin içerikleri bir bütünce (corpus) oluşturmak için kullanılır. Bu bütünce (corpus), dokümanlarda görülen kelimelerin istatistiksel özelliklerini çıkarmak için kullanılır.

Bütünce (corpus) oluşturma işlemi, OCR işlemi sonucunda ortaya çıkan metin dosyasından tüm kelimeleri oluşturmak için kullanılır. Bir dizi işlemden oluşmaktadır. İlk işlem olarak ayırıcı karakterler kullanılarak metin dosyasındaki kelimeler ayrıştırılır. Kelimeleri ayırmak için kullanılan karakterler Çizelge 1'de gösterilmiştir.

Çizelge 1 : Kelimeleri ayırmak için kullanılan karakterler

Karakter	Açıklama
\b	Boşluk
\n	Yeni Satır

Ayrıştırılan kelimeler üzerinde yapılan ilk işlem, kelimelerin küçük harfe çevrilmesidir. Gelen dosyaların Türkçe olduğu bilindiğinde, Türkçe'ye uygun şekilde çevrilme işlemi gerçekleştirilir. Bu adımdan sonra, kelimelerden kaldırılması gereken karakterler kaldırılır. Kelimeden çıkarılan karakter listesi Çizelge 2'de verilmiştir.

Çizelge 2 : Kelimelerden çıkarılan karakterler

Karakter	Açıklama
.	Nokta
,	Virgül
;	Noktalı virgül
:	İki nokta üst üste
*	Yıldız
/	Bölme işareti
!	Ünlem işareti
‘	Tek tırnak
(	Sol parantez
)	Sağ parantez
_	Alt çizgi
\	Ters bölme işareti
~	Tilda

Sonraki işlem, kalan kelimelerden önceden belirlenmiş bazı kelimelerin çıkarılmasıdır. Çıkarılacak kelimeleri içeren liste Çizelge 3'te verilmiştir.

Çizelge 3 : Kelime kümesinden çıkarılacak kelimeler

Kelime
Ve
Veya
Daha
İle
Ki
İse
Vs
De
Da
Dan
Den
Ya
Veveya
Ama
Ki

Son adım olarak, kelimelerden bazı sonekler kaldırılmaktadır. Kaldırılan sonek listesi Çizelge 4'te verilmiştir.

Çizelge 4 : Kelimelerden kaldırılan sonekler

Sonek
Ler
Lar

Bu işlemler sonucu oluşturulan bütüncede kelimeler ve bu kelimelerin eğitimde kullanılan tüm doküman kümesinde kaç defa geçtiği bilgisi bulunur. Bütünce bilgilerini içeren veritabanı tablosundan bir sorgunun görüntüsü Şekil 2'de gösterilmiştir.

	Id	Name	Count
43	238998	tutarı	1463
44	237103	sözleşme	1439
45	236908	sözleşmenin	1339
46	236969	bilgi	1334
47	237000	talep	1326
48	239001	tahsil	1284
49	237081	soyadı	1269
50	237151	yetkili	1267
51	261139	kredinin	1223
52	239147	para	1222
53	257346	bsmv	1133
54	241890	kadar	1124
55	236906	olan	1110
56	237025	işlem	1108
57	236937	imza	1104
58	239167	eden	1093
59	239117	edilen	1084
60	242244	doğan	1076

Şekil 2 : Bütünce bilgilerini içeren veritabanı tablosu

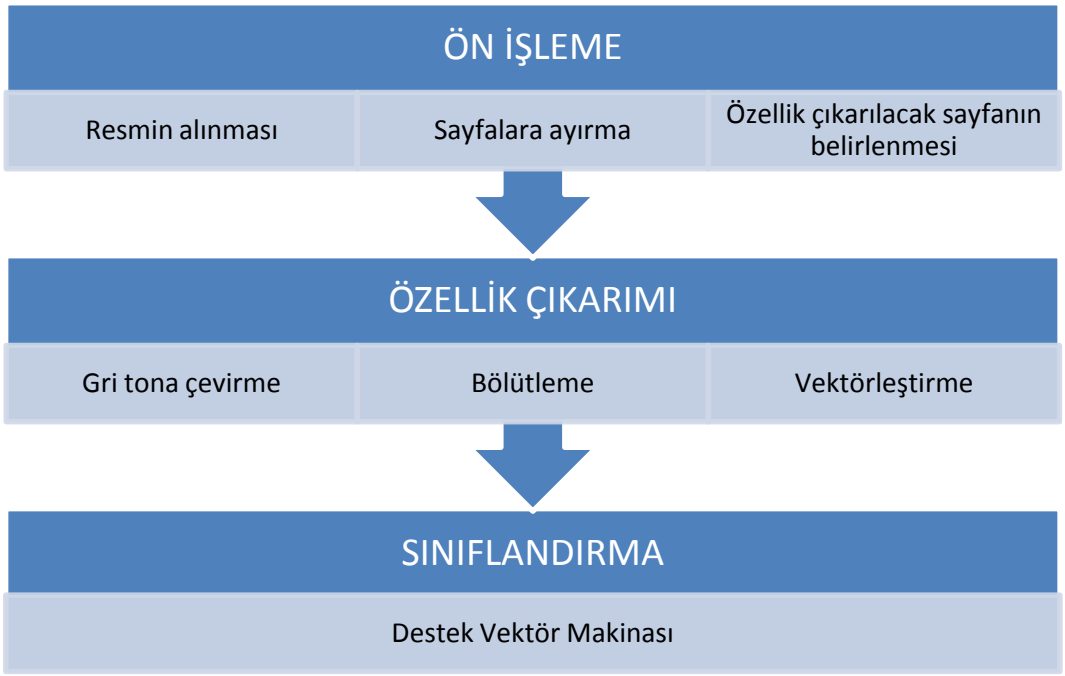
Bütünce oluşturma işleminden sonra yapılacak işlem, öznitelik vektöründe kullanılacak kelimelerin seçilmesi işidir. Bunun için, bütün doküman kümesinde en sık rastlanan kelimeleri temel alan bir öznitelik seçme yöntemi kullanılmıştır. Bir doküman şablonunun parçası olan kelimenin frekans bazlı sıralamada daha yüksek bir sırasının olacağı beklenmektedir. Bunun sebebi, bu kelimenin bu kategorinin her elemanında yer alacak olmasıdır. Bu yüzden ilgili kelime, sınıflandırma için temsil edici bir özellik olabilir. Öte yandan, daha az örnekte (veya sadece bir örnekte) geçen kelimeler, örnek olarak kullanıcıya özel bilgileri içeren kelimeler, dokümanın kategorisi ile ilgili bir bilgi değeri taşımamaktadır. Bu şekilde, kullanıcı ismi nüfus cüzdanı, ehliyet veya bir başvuru formunda görülebilir. Bu terimlerin sıralamada daha düşük değere sahip olacaklarından, öznitelik temsilinde kullanılmazlar.

En sık kullanılan kelimelerin hangi oranda öznitelik temsilinde kullanılacağı önemli bir soru olup, yapılan denemelerle belirlenmeye çalışılmıştır. Gövdelemenin etkisini belirleyebilmek için, basit ancak faydalı bir teknik olan kelimedeki ilk dört harfi alma tekniği kullanılmıştır. Semantik olarak yanlış bir çözüm oluşturabilmekle birlikte, çok verimli ve tatmin eden bir temsil sağlaması beklenmektedir. Bunun sebebi, doküman içeriğinin bu yöntemle sınıflandırma amaçlı özetlenmesidir.

2.2.2 Resim öznitelikleri

Metot

Resim tabanlı doküman sınıflandırma için kullanılan, resim öznitelikleri yaklaşımına ait genel altyapı Şekil 3 ‘de gösterilmiştir. Öznitelik temsili için kullanılacak sayfaların çıkarılmasından sonra, belirlenen sayfalar için öznitelik vektörleri oluşturulmuştur. Sınıflandırma işlemi için ise, destek vektör makineleri kullanılmıştır.



Şekil 3 : Resimden resim tabanlı doküman sınıflandırma için altyapı

Öznitelik çıkarımı

Doküman resmi, resim içeriğini tanımlayan öznitelik kümesi ile temsil edilir. Bu işlemi gerçekleştirmek için, resim alındıktan kaç sayfaya sahip olduğu belirlenir. Eğer birden çok sayfaya sahip ise, doküman resmi sayfalara ayrıştırılır. Bu aşamadan sonra, öznitelik çıkarılacak sayfa belirlenir. Öznitelik çıkarımı için tüm sayfalar kullanılabilir olmakla birlikte, işlemin yüksek performansla gerçekleştirilmesi beklendiğinden tek bir sayfanın öznitelik çıkarılarak doküman kategorizasyon işleminin yapılmasına karar verilmiştir. Eğer doküman çok sayfalı ise, birinci sayfa üzerinde işlem yapılmaktadır.

Öznitelik çıkarımı işlemleri, öznitelik çıkarımı için kullanılan yönteme göre değişmektedir. Gri tona çevirme işlemi ilgili yöntem girdi olarak gri ton bir resme ihtiyaç duyuyor ise gerçekleştirilmektedir. Yöntemlerde sayfanın kendisi bir bütün olarak ele alındığından bölütleme işlemi yapılmamakla birlikte, öznitelik çıkarım yöntemlerinde gereksinim duyulduğunda bölütleme işlemi yapılmaktadır. Sınıflandırma işlemi için ise, destek vektör makineleri yöntemi kullanılmaktadır.

LTP (Local Ternary Patterns) : LTP resimlerdeki dokuyu modellemek için kullanılan bir metottur. Yakın zamanda Suruliandi ve Ramar tarafından LBP (Local Binary Pattern)'ye [31] bir uzantı olarak sunulmuştur [32]. LBP, tekbiçimli olarak terimleştirilen birtakım lokal ikili doku örüntülerini tanımaya dayanmaktadır. Merkezdeki piksel R yarıçaplı olan dairesel bir komşulukta bulunan P pikselleri ile karşılaştırılır. Dairesel çevrenin sınırları boyunca ikili seviyede yapılan bir karşılaştırma bir tekbiçimlilik (uniformity) ölçüsü bulmak için kullanılır. Bu ölçü, geçişleri ortaya koyar. Tekbiçimlilik (uniformity) derecesi önceden tanımlanmış bir eşikten daha az olan bir örüntü, 0 ve P aralığında değişen bir etikete atanır. Tekbiçimli (uniform) olmayan örüntüler ise, tek bir etikete atanır (örnek olarak P+1). LBP öznitelik temsili, ele alınan bölge üzerindeki bu tekbiçimli örüntülerin ayrık oluşma histogramının bir vektöründen oluşur. LTP, ikili bir örüntü üzerinde işleme yerine üçlü bir örüntü üzerinde işlem yapma imkanı sağlar. Geçişlerin sayısını veya örüntülerin dairesel tanımlarındaki süreksizlikleri bulmaya izin verir. Örüntünün tekbiçimliliği, ritmik bir örüntüyü takip ettiği bulunan bu geçişler ile değerlendirilir. Bu örüntülerin geniş bir bölge üzerinde oluşma frekansı LTP öznitelik temsilini oluşturur.

Local ternary pattern geliştirilen altyapıda farklı bir şekilde de kullanılmıştır. Varsayılan yöntemde, kategorize edilmek istenen resmin tamamından tek bir vektör çıkartılmıştır. Kullanılan bir diğer yöntemde ise, resim ortadan dört parçaya bölünmüş ve her bölüm için LTP yöntemi ile oluşturulan öznitelikler ayrı ayrı hesaplanmıştır.

Global öznitelikler : Bu yöntem resmin global temsili ile ilgilidir. Kullanılan global öznitelikler, entropi, en-boy oranı, ve enerji özelliklerini içerir. Verilmiş bir resim bloğu i ve piksel yoğunluğu P_i için, entropi $E_i = -P_i \log P_i$ olarak tanımlanır. Entropi, bu bölgenin içerdiği global bilginin bir ölçüsüdür. Enerji ise, bölütteki tüm piksellerin

yoğunluğunun karelerinin toplamının bölüt alanına bölümünden oluşur. En-boy oranı ise bölütün eninin boyuna bölünmesi ile tanımlanan bir başka global özelliktir.

Bağlı bileşenler tabanlı histogram : Bu yöntemde ilk olarak resimdeki bağlı bileşenlerin etiketlendirilmesi işlemi gerçekleştirilmektedir. Bu işlem ile, öznitelik vektörü elde edilmek istenen resim bölütlere ayrıştırılmış olur. Her bölüt içerdiği siyah piksel sayısına göre bir bölgeye atanarak histogram oluşturulur. Bölütteki siyah piksel sayısının yüzden az olması durumunda, bölüt histograma eklenmez. Burada amaç, gürültü olarak nitelendirilebilecek bölütlerin dikkate alınmak istenmemesidir. Histogram, iki yüz piksel aralıklardan oluşan bölgeleri içermektedir. Histogramda bölgeler iki yüz piksellik artış içerecek şekilde 100-300,300-500,500-700,……,99900-100100 olarak belirlenir. Bölgeler elli piksel kaydırılarak 150-350,350-550,550-750,……..,99950-10150 oluşturulur ve ilk seferde oluşturulan bölgeler eklenir. Böylece histogram bin adet bölgeden oluşturulmuş olur.

Bölütteki siyah piksel sayısı P_b olsun. Bölütün ilk beş yüz bölgedeki yeri aşağıdaki şekilde bulunur :

$$P_1 = (P_b – 100) / 200; \quad (2.1)$$

Eğer bulunan P_1 değeri beş yüzden büyük ise, değeri beş yüz yapılır.

Bölütteki siyah piksel sayısı yüz elliden büyük ise, ikinci beş yüz bölgedeki yeri aşağıdaki şekilde bulunur :

$$P_2 = (P_b – 150) / 200; \quad (2.2)$$

Eğer bulunan P_2 değeri beş yüzden büyük ise, değeri beş yüz yapılır.

Bu şekilde bulunan P_1 ve P_2 değerleri bölütün hangi bölgelere atanacağını belirlemiş olur. İlgili bölgelerdeki sayı arttırılır.

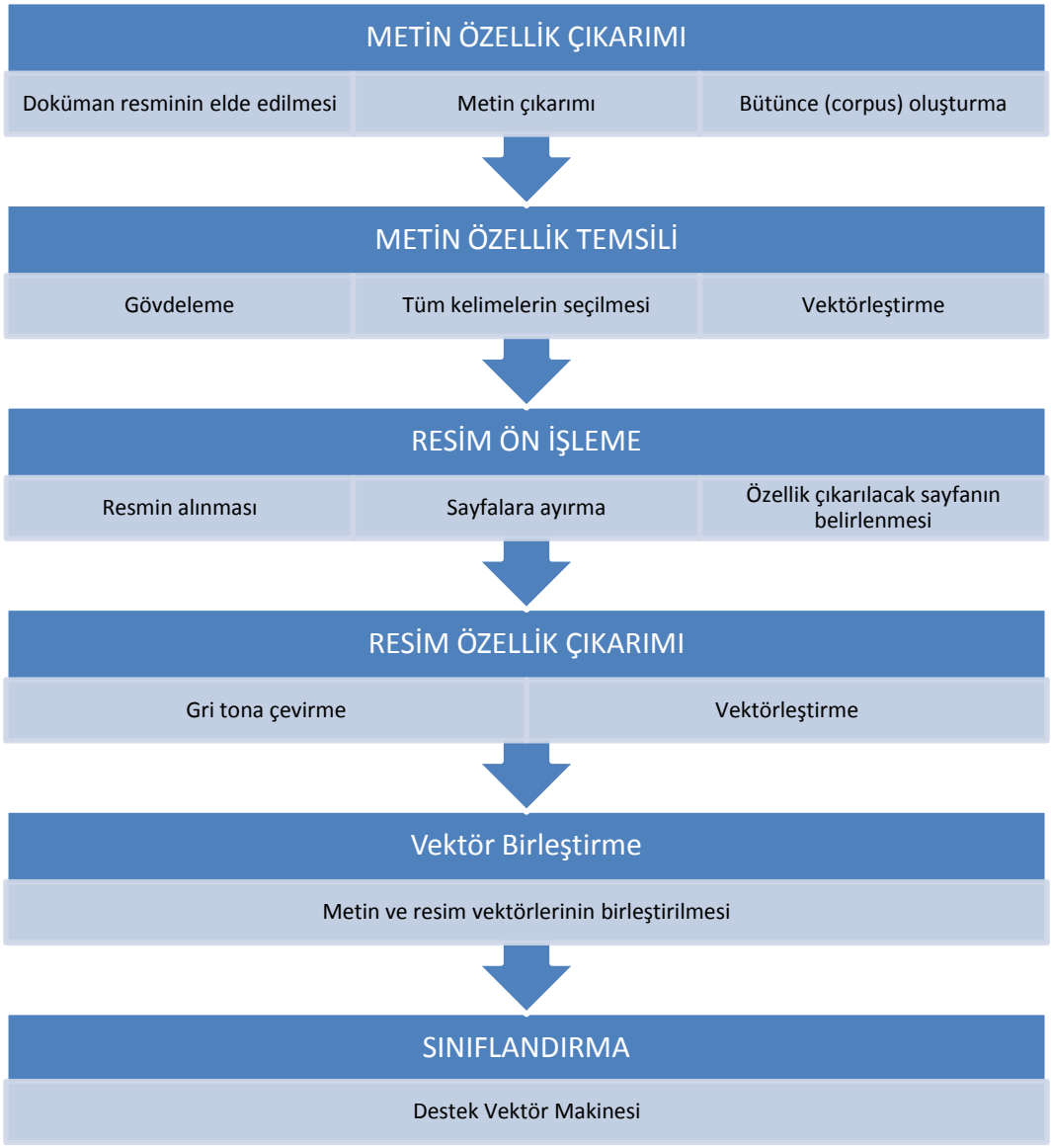
Bu yaklaşım ile, yüz elli pikselden daha fazla siyah piksel içeren bölütler, histogramda tek bir bölgeye değil, iki bölgeye atanmış olurlar. Histogram bölgelerini bu şekilde tasarlamış olmaktaki amaç, bölütün başka bir çizgi vb.

bölütle kesişerek başka bir bölgeye kayması durumunda elli piksel toleransı ile vektörde ikinci temsilde aynı bölgede temsil edilme olasılığı sağlamaktır.

2.2.3 Hibrid yaklaşım

Metot

Hibrid yaklaşımda amaç, metin öznitelikleri ve resim özniteliklerini birleştirerek her birlikte temsil edilmelerinin sağlanmasıdır. Metin ve resim özniteliklerinde en iyi doğruluğa sahip yöntemlerin birlikte kullanılmıştır. Buna göre metin özniteliklerinde en iyi doğruluk oranı, tüm kelimeler kullanıldığında ve gövdeleme işlemi yapıldığında elde edilmiştir. Resim özniteliklerinde en iyi doğruluk oranı ise, local ternary patterns ve global öznitelikler birlikte kullanıldığında elde edilmiştir. Buna göre hibrid yöntem yöntemin çalışma Şekil 4'te gösterilmiştir.



Şekil 4 : Hibrid yöntemle doküman sınıflandırma için altyapı

Öznitelik çıkarımı ve temsili

Metin öznitelikleri ve resim öznitelikleri birlikte kullanılarak bir öznitelik vektörü oluşturulduğundan, her iki işlem için belirlenen yöntemler sırayla çalıştırılmaktadır. Bazı alt adımların da bu durumda netleştirilmesi gerektiğinden yöntem sırayla açıklanmıştır.

Doküman resminin elde edilmesi : Kategorize edilmek istenen dokümana ait resim dosyası uygulamaya verilir. Doküman çok sayfalı ise TIFF, diğer durumda jpg formatında oluşturulmuştur. Uygulama dokümanın formatı ile ilgili codec dosyasını kullanarak dokümanı bitmap haline çevirir.

Metin çıkarımı : Resim formatındaki dosyadan metin dosyası elde etmek için OCR işlemi uygulanmaktadır. Hibrid yöntemde Abby OCR motoru kullanılmıştır. OCR işlemi sonucu oluşan metin dosyasında Türkçe karakterler mevcuttur.

Bütünce (corpus) oluşturma : İşlem metin öznitelikleri bölümünde belirtildiği şekilde gerçekleştirilir. İşlemden sonra oluşan kelimeler, veritabanında bir tabloda saklanır.

Gövdeleme : Kelimenin ilk dört harfi alınır.

Tüm kelimelerin seçilmesi : Bulunan tüm kelimeler, öznitelik vektöründe kullanılmak üzere seçilir.

Vektörleştirme : Seçilen doküman için, öznitelik vektöründe geçen kelimelerin kaç defa geçtiğini içeren vektör oluşturulur.

Resmin alınması : Doküman resminin elde edilmesi ile aynı adımları içerir.

Sayfalara ayırma : Resim tabanlı doküman kategorizasyonunda doküman sayfalara ayrıştırılır. Bu işlem, doküman çok sayfa içeren tiff formatında oluşturulmuş ise uygulanır.

Öznitelik çıkarılacak sayfanın belirlenmesi : Önceki adımda ayrıştırılan sayfalardan ilk sayfa seçilir.

Gri tona çevirme : LTP yönteminin uygulanabilmesi gelen dokümanın gri tona çevrilmesi gerektiğinden, bu adımda doküman gri tona çevrilir.

Vektörleştirme : LTP yöntemi tüm sayfaya uygulanır. Bu işlemden 512 tamsayı üretilir. Resim ortadan dört bölgeye bölünerek her bölgeye LTP yöntemi uygulanır. Son olarak global öznitelikler vektöre eklenir. Öznitelik vektöründe toplam 2563 tamsayı bulunmaktadır.

Vektör birleştirme : Metin özniteliklerinden oluşturulan vektör ile resim özniteliklerinden oluşturulan vektör yan yana eklenir.

2.2.4 Sınıflandırma

Multinom Naive Bayes

Doküman resim sınıflandırma için kullanılacak ilk sınıflandırıcı Naive Bayes sınıflandırıcısıdır. Naive Bayes sınıflandırıcı Bayes istatistiğine dayanan bir öğreticiyle öğrenme (supervised learning) sınıflandırma tekniğidir. Bayes istatistiği veri seti altında yatan olasılıksal bir modelin varlığını varsayar. Bu model ortaya çıkabilecek sonuçlara göre model hakkında belirsizliği ortaya koyar.

Problemdeki öznitelik vektörü elemanlarının gerçekte birbiri ile ilişkisi olmasına rağmen bu elemanların birbirinden bağımsız olduğunu kabul eder. Böylece her elemanın problemin çözümüne diğer elemanlardan bağımsız olarak katkı sağladığını farz eder. İsmindeki Naive sıfatını bu kabulden dolayı almıştır. Yöntemdeki bu "naive" varsayıma rağmen, çoğu gerçek hayat problemlerinde performansı yüksektir.

Naive bayes modellerde parametre tahminini en yüksek olasılık (maximum likehood) kullanılarak yapılır. Multinom Naive Bayes (MNB) yönteminde, multinom olasılık dağılımı olduğu kabul edilmiştir.

$X = \{x_1, x_2, \dots\dots, x_d\}$ değişkenleri verildiğinde, olası $C = \{c_1, c_2, \dots, c_d\}$ sınıfları arasından C_j sınıfının sonsal olasılığı (posterior probability) oluşturulmak istenmektedir. Daha bilinen bir ifadeyle, X öznitelik vektörü, C ise sınıflar kümesidir. Bayes kuralı kullanılarak yazılan aşağıdaki ifadede:

$$\left(C_j \middle| x_1, x_2, \dots., x_d\right) = p\left(x_1, x_2, \dots., x_d \middle| C_j\right) p(C_j) \tag{2.3}$$

$p\left(C_j \middle| x_1, x_2, \dots., x_d\right)$ sınıfa ait olma ile ilgili sonsal olasılıktır, bu da X in C_j sınıfına ait olma olasılığıdır. Naive Bayes bağımsız değişkenlerin koşullu olasılıklarını istatistiksel olarak bağımsız kabul ettiğinden, ihtimali terimlerin çarpımına çevirebiliriz:

$$p\left(X \middle| C_j\right) = \prod_{k=1}^{d} p\left(x_k \middle| C_j\right) \tag{2.4}$$

Bu durumda sonsal olasılık aşağıdaki şekilde tekrar yazılabilir:

$$p(C_j|X) = p(C_j) \prod_{k=1}^{d} p(x_k|C_j) \quad (2.5)$$

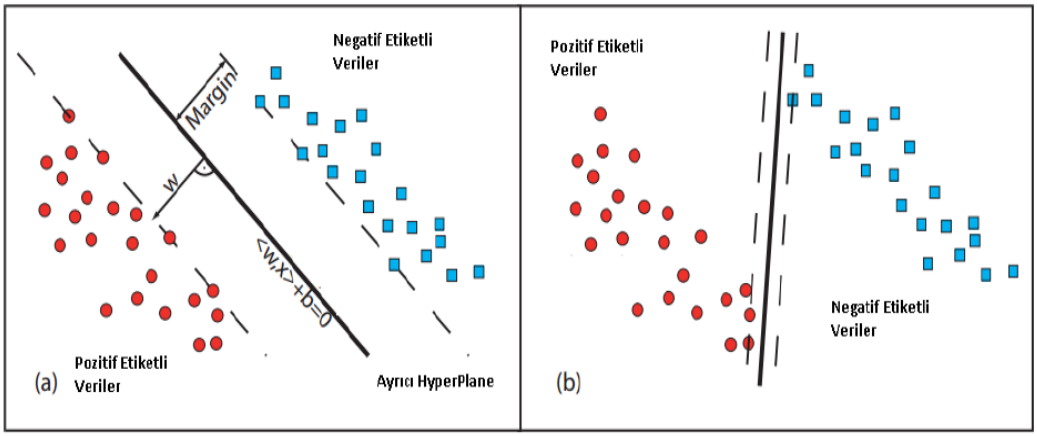
Yukarıdaki Bayes kuralını kullanarak, bir X örneğini en yüksek sonsal olasılığa ulaşan C_j sınıfı ile etiketleyebiliriz. Değişkenlerin bağımsız olması varsayımı her zaman doğru olmamakla birlikte, sınıflandırma işlemini büyük oranda kolaylaştırmaktadır. Bunun sebebi $p(x_k|C_j)$ koşullu olasılıklarının her değişken için ayrı bir şekilde hesaplanmasına izin vermesidir. Böylece çok boyutlu iş, birçok tek boyutlu işe indirgenmiş olur. Bunun ötesinde, varsayım sonsal olasılıkları büyük ölçüde değiştirmediğinden, sınıflandırma işini etkilemez.

Destek vektör makineleri

Destek vektör makineleri iki sınıf arasında tahmin yapan bir sınıflandırıcıdır. Bu sınıflandırıcı yapısal riskleri en aza indirme prensibine göre çalışır. SVM n boyutlu girdi verisini doğrusal olmayan bir şekilde daha yüksek bir boyuta taşır. Taşıdığı bu yüksek boyutta doğrusal bir sınıflandırıcı oluşturur. SVM yöntemi sınıflandırma aşamasına geldiği zaman, sınıflandırılacak dizilim için eğitim safhasındaki gibi dizilimi temsil eden bir öznitelik vektörüne ihtiyaç duyar. Bu öznitelik vektörü test veri setindeki her bir veri için ayrı ayrı oluşturulmalıdır.

Tanım olarak SVM verilen iki veri kümesi arasında veriyi bir birinden ayıran optimum hiper düzlemi bulur. Aşağıda Şekil 5b'de veri kümesi iki boyutlu olduğu için verileri ayıran hiper düzlem çizgidir. Verileri ayıran bir çok çizgi olmasına rağmen şekil 5a'da bulunan ayırım göz ile de fark edilebileceği gibi en optimum ayırımdır.

Yine şekilde görüldüğü gibi veriyi birbirinden ayıran birçok düzlem olmasına rağmen SVM'nin amacı Şekil 6'da görüldüğü gibi maksimum margin ile ayırım yaparak destek vektörlerini bulmaktır.



Şekil 5 : İki farklı hiper düzlem

SVM çıktısı test edilecek veriye ait ayırt edici skorudur. İki sınıf arasında sınıflandırma yapan sınıflandırıcılarda pozitif skor verinin o sınıfa ait olduğuna işaret eder. Sistemde sıfırdan büyük değerler iyi bir skor olarak kabul edilmiştir.

Tez çalışmasının doküman kategorizasyon kısmında, SVM doğrusal, polinom ve gauss dağılımlı radyal çekirdekleri ve LIBSVM [14] uygulamasındaki varsayılan girdi parametreleri ile kullanılmıştır.

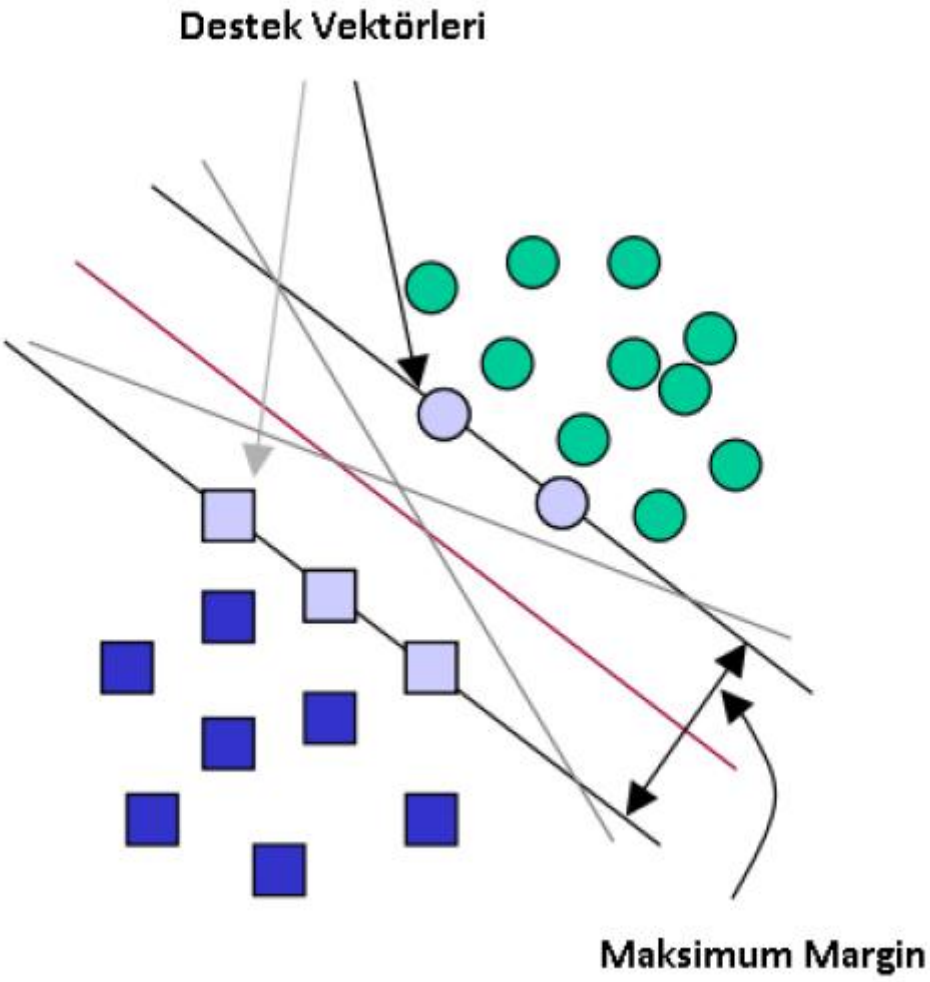
Matematiksel olarak SVM aşağıdaki gibi tanımlanır.

$\{x_i, y_i\}$, i=1,….,N, olarak verilen eğitim setinde her örnek d tane özelliğe sahiptir ($x_i \in R^d$). Sınıfları belirten y_i sadece iki değer alabilir ($y_i \in \{1,-1\}$). D boyutlu bu uzayda bütün hiper düzlemler bir vektör ve bir sabit sayı ile belirtilir. Bu ifadeyi aşağıdaki gibi belirtebiliriz [33].

$$w * x + b = 0 \tag{2.6}$$

Not olarak w vektörü hiper düzleme dik bir vektördür. Bu formülü SVM nin sınıflandırma fonksiyonu olarak kullanırsak aşağıdaki formüle ulaşırız.

$$f(x) = sign(w * x + b) \tag{2.7}$$

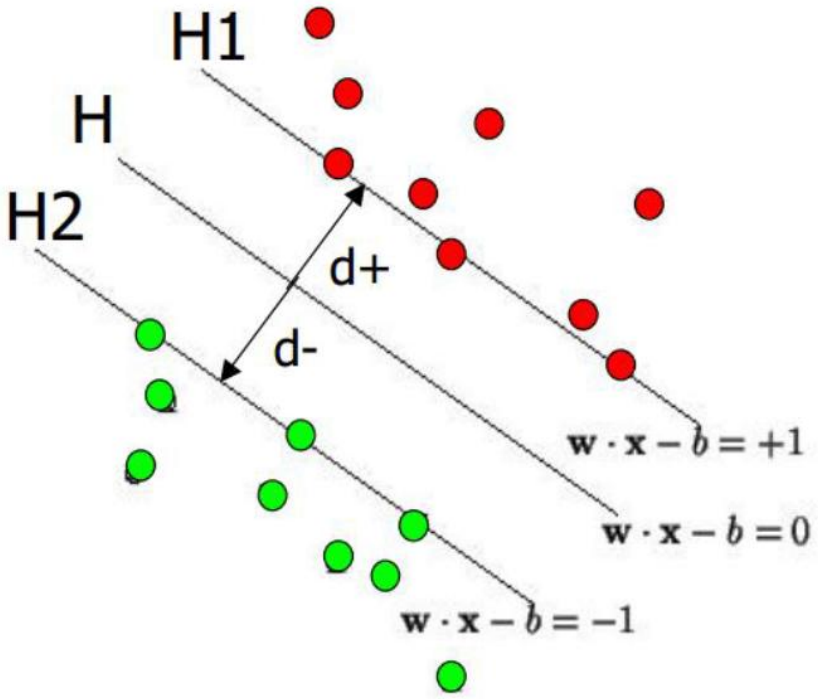


Şekil 6 : Destek vektörleri

Veri setinde bulunan herhangi örnek olan x değeri formülde yerine koyulursa şekil 6'da görüldüğü üzere aşağıdaki gibi bir sonuç ortaya çıkar.

$$x_i * w + b \geq +1 \; ise \; y_i = +1 \tag{2.8}$$

$$x_i * w + b \leq -1 \; ise \; y_i = -1 \tag{2.9}$$



Şekil 7 : Geometrik olarak hiper düzlem

Veya daha sade olarak ;

$$y_i(x_i * w + b) \geq 1 \tag{2.10}$$

ifadesi veri setindeki her örnek içi doğru olur.

Geometrik olarak x_i noktasının hiper düzleme olan uzaklığı hesaplanırken w nin değeri normalize edilmelidir. Böylece x_i noktasının hiper düzleme olan uzaklığı basitçe aşağıdaki gibi formül haline getirilebilir.

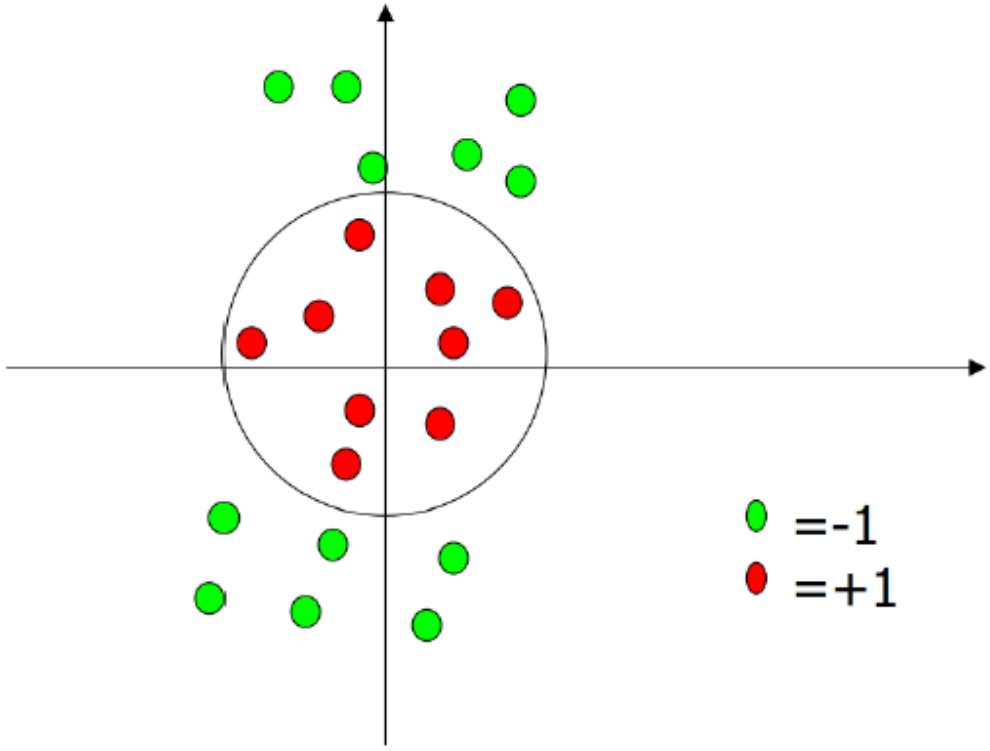
$$d\big((w,b),x_i\big) = \frac{y_i(x_i * w+b)}{\|w\|} \geq \frac{1}{\|w\|} \tag{2.11}$$

Bu noktanın hiper düzleme olan uzaklığını maksimize edilmek istendiği için yukarıdaki formüldeki ||w|| ifadesi minimize edilmesi gerekmektedir. Bu ifadenin minimize edilmesinde kullanılan başlıca yöntem Vapnik de belirtildiği gibi Lagrange çarpanlarıdır [34]. Bu yöntem kullanılarak ifade aşağıdaki ifadenin minimize edilmesine dönüştürülür.

$$W(\alpha) = -\sum_{i=1}^{N} \alpha_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j (x_i^T * x_j) \tag{2.12}$$

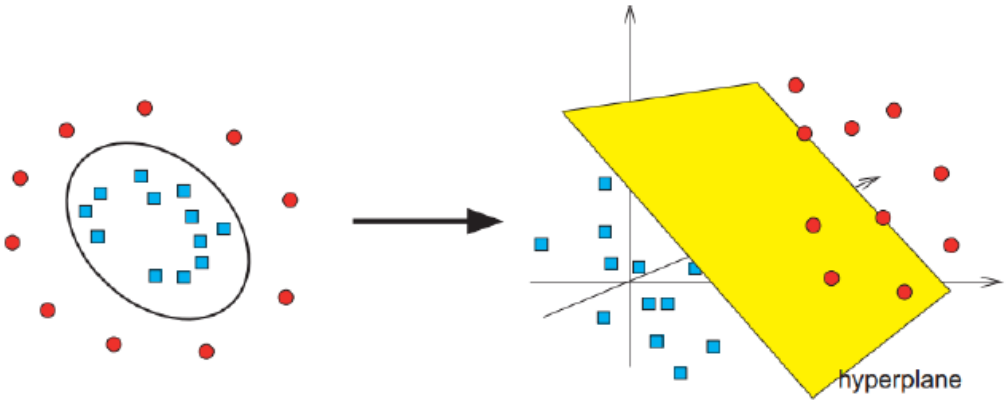
Bu ifadenin minimize edilmesi ile her veri için bir tane olmak üzere toplam L tane α değeri bulunur [39]. Bulunan alfa değerlerinden sıfırdan büyük olanlar destek vektörleri olarak tanımlanmıştır. Örnek olarak 1000 verilik bir eğitim setinde çıkan α değerlerinin birçoğu sıfır olacaktır [39]. Bu noktalar veriyi ayıran maksimum margin ile tanımlamış hiper düzlemin dışında kalan noktalardır. Fakat α_i değeri sıfırdan büyük ise bu değerin ait olduğu x_i vektörü destek vektörü olarak tanımlanır. Destek vektörlerinin bulunması ile doğrusal olarak ayrılan veriler için maksimum margine sahip hiper düzlem bulunmuş olur.

Şu ana kadar veri setinin doğrusal olarak ayırabilindiği farz edilmiştir. Ama karşılaşılan problemlerin birçoğunda veri setinde doğrusal olarak ayrılamaz. Bu durum Şekil 8'de örnek olarak gösterilmiştir.



Şekil 8 : Doğrusal olarak ayrılamayan veri seti

Uygun bir Φ fonksiyonu ile veri setinin doğrusal olarak ayrılabileceği yüksek boyutlu bir sisteme taşındığı farz edilirse, yeni oluşan çok boyutlu uzay özellik uzayı H olarak adlandırabilir. Bu uzayda bulunan bir hiper düzlem ile mevcut veriler doğrusal olarak ayrılacaktır [40] (Şekil 9).



Şekil 9 : Veri setinin hiper düzlemde doğrusal olarak ayrılması

Doğrusal olarak ayrılamayan veriler için ulaşılan optimum hiper düzlemin formülü, doğrusal olarak ayrılabilen veriler için olan ile birebir aynıdır. Tek fark formüldeki x_i vektörlerinin d boyut olması yerine, $\Phi(x_i)$ vektörünün daha yüksek belki de sonsuz boyutta olmasıdır.

$$W(\alpha) = -\sum_{i=1}^{N} \alpha_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j (\phi(x_i^T) * \phi(x_j)) \qquad (2.13)$$

Formül incelendiği zaman görülen en önemli nokta yüksek boyutlu uzaydaki vektörlerin nokta çarpımları ile ilgilidir. Elimizdeki vektörlerin yüksek boyutlu uzaya taşınmış halindeki $\phi(x_i^T) * \phi(x_j)$ nokta çarpımını yüksek boyutlu uzayda yapılması çok maliyetli bir işlemdir [39]. Hatta bazı durumlarda veri sonsuz boyutlu uzaya taşındığı için bu işlemi yapmak gerçek manada imkânsızdır. Bu noktada çekirdek fonksiyonları verinin transfer edilmiş uzaydaki nokta çarpımlarını verirler. Çekirdek fonksiyonları yardımı ile verinin transfer edildiği yüksek boyutlu uzay hakkında hiçbir şey bilinmese bile bu uzaylar kullanabilir [39]. Bu durum aşağıdaki gibi formülleştirilmiştir.

K çekirdek fonksiyonu ve Φ vektörleri yüksek boyuta taşıma fonksiyonu olmak üzere

$$K(x, y) = \phi(x) * \phi(y) \qquad (2,14)$$

Bu durumun direk bir sonucu olarak vektörleri yüksek boyuta taşıyan fonksiyon hakkında hiçbir şey bilinmese bile çekirdek fonksiyonları ile destek vektör makineleri verimli bir şekilde kullanılabilir [35]. Çekirdek fonksiyonları destek vektör

makinelerinin en önemli ve anlaşılması zor konusudur [36]. Hangi çekirdek fonksiyonun seçileceği probleme bağlı olarak değişir. Doğru çekirdek bulunsa bile çekirdek parametrelerini seçmek zor olabilir. Çekirdek seçme işini otomatik olarak yapma konusunda çeşitli çalışmalar yapılmıştır [37]. Altyapıda denenmiş olan çekirdek fonksiyonları aşağıda incelenecektir.

Polinom çekirdek, eğitim veri setinde ki bütün değerlerin normalize edildiği problemler için iyi bir seçimdir. Aşağıdaki gibi formülleştirilebilir [38].

$$K(x, y) = (\alpha x^T y + c)^{\,d}, \alpha > 0 \tag{2.15}$$

Doğrusal çekirdek fonksiyonları, vektörlerin iç çarpımlarına sabit bir değer ekleyerek bulunur. Verinin doğrusal olarak ayrılamadığı durumlarda kullanılması iyi bir seçim değildir. Aşağıdaki gibi formülleştirilebilir [38].

$$K(x, y) = x^T * y + c \tag{2.16}$$

Radyal temelli çekirdek, datayı doğrusal olmayan bir şekilde yüksek boyuta taşır. Doğrusal çekirdeğin aksine datanın doğrusal olarak ayrılamadığı koşullarda verimli bir şekilde çalışabilir. Öznitelik vektörünün sayısının çok yüksek olduğu durumlarda kullanılması tavsiye edilmez, aşağıdaki gibi formülleştirilebilir [38].

$$K(x, y) = \exp\left(-\alpha \left\| x_i - x_j \right\|^2\right), \alpha > 0 \tag{2.17}$$

2.3 Sonuçlar

Özel bir gizlilik anlaşması ile bir Türk bankasından on dokuz doküman grubuna ait gerçek müşteri dokümanları toplanmıştır. Çizelge 5'de doküman kategorileri ve her kategorideki resim formatı, renk derinliği, sayfa sayısı, ve ilgili kategorideki doküman sayısı listelenmiştir. Her kategori için dokümanların yarısı eğitim kümesinde, diğer yarısı test kümesinde kullanılmıştır.

Bazı kategorilere ait dokümanların farklı sayfa adedi içerdiği görülmektedir. Örnek olarak ilk doküman grubu olan bireysel bankacılık hizmet sözleşmesi dokümanı her zaman bir sayfadan oluşurken, müşteri talimatları farklı sayfa adedi içeren dokümanlardan oluşabilmektedir. Aynı doküman grubunda en yüksek sayfa adedi farkı, ödeme planı ve kredi şartları doküman grubunda oluşmuştur.

Renk derinliği incelendiğinde, dokümanların genelde siyah/beyaz olduğu, sadece üç doküman grubunun daha yüksek sayıda renk içerdiği görülmektedir. Bunlardan ehliyet doküman grubu her zaman renkli olmakla birlikte, çek ve nüfus cüzdanı doküman grupları hem siyah/beyaz hem de renkli olabilmektedir.

Çalışmada, farklı kategorilerdeki örnek sayısı birbirine yakın olduğundan, modellerin kategorizasyon performansını değerlendirmek için, doğruluk (accuracy) metriği kullanılmaktadır. Doğruluk basit bir şekilde, doğru sınıflandırılmış örneklerin öngörülme yapılan bütün test kümesi içinde yüzdesi olarak tanımlanabilir.

Çizelge 5 : Oluşturulan veri kümesinde bulunan doküman kategorileri

No	Doküman Kategorisi	Doküman Sayısı	Sayfa Sayısı	Format	Renk Derinliği
1	Bireysel Bankacılık Hizmet Sözleşmesi	50	1	Tiff	1
2	Bireysel Ürünler Başvuru Formu	50	2	Tiff	1
3	Çek	99	2/ 1	Tiff/Jpg	1/ 24
4	Havale Eft Formu	50	1	Tiff	1
5	Kart Teslim Belgesi	99	1	Tiff	1
6	Kimlik Fotokopisi	50	1	Tiff	1
7	Kredi Sözleşmesi	50	11	Tiff	1
8	Müşteri Talimatı	99	1-2-3	Tiff	1
9	Nüfus Cüzdanı	50	1	Jpg/ Tiff	24/ 1
10	Ehliyet	50	1	Jpg	24
11	Firma İmza Sirküleri	46	1-3-4-5-7	Tiff	1
12	Hayat Poliçesi	108	2-3-4-5	Tiff	1
13	Kredi Şartları	6	2-12	Tiff	1
14	Kredi Talep Formu	50	1-2	Tiff	1
15	Ödeme Planı ve Kredi Şartları	70	3-4-5-6-22	Tiff	1
16	Sermaye Piyasası İşlem Risk Bildirim Formu	31	1-2	Tiff	1
17	Sigorta Poliçesi	6	2	Tiff	1
18	Temel Bankacılık Ürün Bilgi Formu	50	3-5	Tiff	1
19	Yerleşim Belgesi	50	1	Tiff	1

2.3.1 Metin öznitelikleri

Bütüncede (corpus) bulunan kelimeler incelendiğinde, kullanılan OCR motoruna göre, bütüncenin içerdiği kelime sayısının değiştiği görülmüştür. Gövdeleme işlemi yapılmadan, Tesseract motoru kullanılarak oluşturulan bütüncede bulunan eşsiz kelime sayısı 90890'dır. Aynı işlem Abby OCR motoru kullanılarak yapıldığında bütüncede bulunan eşsiz kelime sayısı 49571 olmaktadır.

Kelimeler bütün kümedeki belirme sayılarına göre sıralandırılmaktadır. Öznitelik seçme stratejisi, bütün dokümanlar kümesinde en sık tekrarlanan kelimeleri öznitelik vektörüne ekleme şeklinde uygulanmaktadır. Seçilen kelime sayısı 100 ile 15000 arasında değişmektedir. Seçim hem gövdeleme yapılarak hem de gövdeleme yapılmadan uygulanmaktadır.

Destek vektör makineleri ve MNB, Weka yazılımı ile kullanılmıştır. MNB yazılımı Weka yazılımı içinde bulunmaktadır. Destek vektör makinelerini Weka yazılımı içerisinden kullanabilmek için LIBSVM [14] paketi kullanılmıştır.

Destek vektör makineleri LIBSVM paketi içinde bulunan varsayılan parametreleri ile uygulanmaktadır. MNB'yi test etmek için, Weka yazılımı içinde bulunan MNB yazılımı varsayılan parametreleri ile kullanılmaktadır. Modeller hem eğitim hem de bağımsız test kümesinde çalıştırılmaktadır. Yapılan denemeler ile ilgili sonuçlar Çizelge 6'da gösterilmiştir.

Çizelge 6 : Metin çıkarımı tabanlı yöntemler ile sınıflandırma sonuçları

Metot			**Doğruluk (%)**			
Seçilen Kelime Sayısı	Gövdeleme	Sınıflandırma Yöntemi	Tesseract OCR Motoru		Abby OCR Motoru	
			Eğitim Seti	Test Seti	Eğitim Seti	Test Seti
100	hayır	SVM	89,66	77,33	89,56	75,04
		MNB	72,74	69,49	70,67	66,16
	evet	SVM	91,63	78,94	93,79	81,66
		MNB	76,50	71,76	74,15	72,21
200	hayır	SVM	90,88	77,24	90,88	76,08
		MNB	74,81	70,53	74,15	67,58
	evet	SVM	93,89	79,32	95,77	81,47
		MNB	79,51	74,03	79,13	72,96
500	hayır	SVM	94,92	78,47	95,11	80,81
		MNB	81,57	74,40	82,04	75,89
	evet	SVM	96,89	79,88	97,36	82,13
		MNB	82,33	75,54	83,74	78,44
1000	hayır	SVM	96,80	78,56	96,80	80,62
		MNB	84,39	74,88	85,71	80,15
	evet	SVM	98,40	79,69	98,12	82,79
		MNB	85,71	77,05	86,74	79,58
2500	hayır	SVM	98,21	79,03	98,02	81,19
		MNB	88,90	80,17	86,46	80,81
	evet	SVM	99,15	80,73	98,96	84,49
		MNB	91,82	79,22	89,94	81,56
5000	hayır	SVM	99,53	81,01	98,87	72,96
		MNB	93,04	82,15	89,94	68,43
	evet	SVM	99,62	79,88	99,06	84,21
		MNB	93,60	81,39	91,63	82,32
10000	hayır	SVM	99,62	79,41	99,15	82,23
		MNB	94,17	82,24	91,91	81,38
	evet	SVM	99,62	81,39	99,24	84,12
		MNB	94,92	81,96	92,38	82,41
15000	hayır	SVM	99,62	80,35	99,62	79,32
		MNB	94,64	82,05	95,30	81,96
	evet	SVM	99,24	82,23	99,24	83,83
		MNB	92,95	81,47	84,11	81,85
Hepsi	hayır	SVM	99,90	66,38	99,24	81,75
		MNB	96,42	38,99	93,23	76,37
	evet	SVM	99,90	77,62	99,24	85,57
		MNB	95,39	76,39	92,57	89,42

Birbirlerine yakın olmakla birlikte, SVM genelde MNB'ye göre daha iyi performans vermektedir.

SVM çekirdek fonksiyonları arasında en iyi sonuç doğrusal çekirdek fonksiyonu kullanılarak elde edilmiştir.

Gövdeleme işlemi, doğruluğu arttırmaktadır.

Tüm sonuçlar değerlendirildiğinde, test kümesindeki en yüksek doğruluk (%89,42) olup, MNB sınıflandırıcısı ile tüm kelimeler kullanıldığında ve gövdeleme işlemi yapıldığında elde edilmiştir.

Kelimelerin tamamı nitelik vektöründe temsil edildiğinde, kelime sayısı kullanılan OCR motoruna göre değişkenlik göstermektedir. Tesseract OCR motoru kullanıldığında 90989 kelime olup, ilk dört harf alındığında farklı kelime sayısı 49571 olmaktadır. Abby OCR motoru kullanıldığında ise, farklı kelime sayısı 73801, ilk dört alınınca oluşan kelime sayısı 41490 olarak ortaya çıkmaktadır.

2.3.2 Resim öznitelikleri

Resim öznitelikleri çıkarımı için kullanılan yöntemlere göre, farklı sayıda tamsayı içeren öznitelik vektörleri oluşmuştur. Metotlarda kullanılan öznitelik çıkarım yöntemleri ve uygulama şekilleri değişkenlik göstermektedir.

İlk metotta, gelen doküman resmi ortadan dörde bölünmüş ve her bölüme LTP (Local Ternary Patterns) öznitelik çıkarımı uygulanmıştır. Sonraki adımda, resmin bütününden global öznitelikler hesaplanmıştır.

İkinci metotta, gelen resmin bütününe LTP yöntemi uygulanmıştır. İkinci adımda, resmin bütününden global öznitelikler hesaplanmıştır.

Üçüncü metotta, resmin tamamına bağlı bileşenlerin histogramı yöntemi uygulanmıştır. Sonraki adımda, resmin bütününden global öznitelikler hesaplanmıştır.

Dördüncü metotta, gelen doküman resmi dörde bölünerek her bölüme LTP öznitelik çıkarımı uygulanmıştır. İkinci adımda, resmin tamamına LTP yöntemi, son adımda da resmin tamamına global öznitelikler yöntemi uygulanarak öznitelik vektörü oluşturulmuştur.

Beşinci metotta, gelen doküman resmi dörde bölünerek her bölüme LTP öznitelik çıkarımı uygulanmıştır. İkinci adımda, resmin tamamında LTP yöntemi, üçüncü adımda da resmin tamamında global öznitelikler yöntemi kullanılmıştır. Son adımda resmin tamamına bağlı bileşenlerin histogramı yöntemi uygulanarak öznitelik vektörü oluşturulmuştur.

Metotlarda kullanılan öznitelik çıkarım yöntemleri, açıklama ve öznitelik vektöründe kullanılan öznitelik sayısı Çizelge 7'de gösterilmiştir.

Çizelge 7 : Resim öznitelikleri tabanlı metotlar

Metot	Öznitelik Çıkarım	Resme Uygulanma Şekli	Vektördeki Eleman Sayısı
1	LTP	4 x ¼	2051
	Global Öznitelikler	Resmin tamamına	
2	LTP	Resmin tamamına	515
	Global Öznitelikler		
3	Bağlı Bileşenlerin Histogramı	Resmin tamamına	1003
	Global Öznitelikler		
4	LTP	4 x ¼ + Resmin tamamına	2563
	Global Öznitelikler	Resmin tamamına	
5	LTP	4 x ¼ + Resmin tamamına	3563
	Global Öznitelikler	Resmin tamamına	
	Bağlı Bileşenlerin Histogramı		

Eğitim ve test işlemleri Weka yazılımı kullanılarak yapılmıştır. Destek vektör makinelerini Weka yazılımı içerisinden kullanabilmek için LIBSVM [14] paketi kullanılmıştır.

Destek vektör makineleri LIBSVM paketi içinde bulunan varsayılan parametreleri ile uygulanmaktadır. Destek vektör makineleri paket içinde bulunan tüm çekirdek fonksiyonları ile denenmiş, en iyi sonuç üreten çekirdek fonksiyonu ile elde edilen değer raporlanmıştır.

Modeller hem eğitim hem de bağımsız test kümesinde çalıştırılmaktadır. Yapılan denemeler ile ilgili sonuçlar Çizelge 8'de gösterilmiştir.

Çizelge 8 : Resim çıkarımı tabanlı yöntemler ile sınıflandırma sonuçları

Metot	**Öznitelik Çıkarımı**	**Doğruluk (Accuracy)**	
		Eğitim Seti	**Test Seti**
1	LTP Global Öznitelikler	100	69,78
2	LTP Global Öznitelikler	96,89	68,93
3	Bağlı Bileşenlerin Histogramı Global Öznitelikler	95,10	67,23
4	LTP Global Öznitelikler	100	69,87
5	LTP Global Öznitelikler Bağlı Bileşenlerin Histogramı	99,24	67,32

Sonuçlara göre, kullanılan yöntemler birbirlerine yakın doğruluğa sahiptir. En iyi sonuç, %69,87 ile LTP ve global özniteliklerin birlikte kullanıldığı dördüncü metot ile elde edilmiştir.

2.3.3 Hibrid yaklaşım

Hibrid yaklaşımda, metin öznitelikleri ve resim özniteliklerini birleştirerek hep birlikte temsil edilmelerinin sağlanmak istendiğinden, metin ve resim özniteliklerinde en iyi doğruluğa sahip yöntemler birlikte kullanılmıştır.

Öznitelik vektöründe oluşan eleman sayısı, bu iki yöntemde kullanılan eleman sayılarının toplamından oluşmaktadır. Birleştirilmiş öznitelik vektöründe 44053 eleman bulunmaktadır.

Metin özniteliklerinde en iyi sonuç veren yaklaşımda, Abby OCR motoru ile oluşturulan metinde 73801 farklı kelime üretilmektedir. İlk dört harf dikkate alındığında, eşsiz (unique) kelime sayısı 41490 olmaktadır. Bu kapsamda metin özniteliklerinden gelen vektördeki öznitelik sayısı 41490'dır. Resim özniteliklerinde en iyi sonucu veren metotta, 2563 elemandan oluşan vektör kullanılmaktadır. Her iki vektör birleştirildiğinde, toplamda 44053 elemandan oluşan vektör ortaya çıkmaktadır.

Hibrid öznitelikleri içeren metot kullanıldığında, yapılan denemeler ile ilgili sonuçlar Çizelge 9'da gösterilmiştir.

Çizelge 9 : Hibrid yöntem ile sınıflandırma sonuçları

Sınıflandırıcı	Doğruluk	
	Eğitim Seti	Test Seti
SVM - Polynom	100	69,94
SVM - Doğrusal	99,06	70,60
MNB	31,67	27,41

Sonuçlara göre SVM, MNB yöntemine göre daha yüksek doğruluğa sahiptir. En yüksek doğruluk, doğrusal çekirdek fonksiyonu kullanan SVM ile elde edilmiştir.

3. İMZA BÖLGE ANALİZİ

3.1 Giriş

Bankacılık uygulamalarında elektronik verinin kullanımının giderek artmasına rağmen, imza, basılmış dokümanlarda kullanıcının kimliğinin belirlenmesi ve onayının alınması için en güvenli yöntem olarak değerlendirilmektedir. Banka çalışanları dokümanın imzalı kopyasını taramakta ve imza doğrulama işlemini kendileri gerçekleştirmektedirler. Kullanıcının kendisinin yaptığı bu kontrol, hataya açık olmakla birlikte ek bir iş yükü oluşturmaktadır. Bu durum, banka iş süreçlerindeki maliyeti arttırmakta ve mevcut personel gücü ile daha yoğun çalışmayı gerektirmektedir. Bu durum sigorta sektörü ve diğer sektörler için de geçerliliğini korumaktadır. Bu yüzden, imza ile ilgili işlerin gerçekleştirilmesinde, dokümanların otomatik analizini sağlayan zeki yazılım tekniklerinin kullanımı gerekli hale gelmiştir.

Taranan başvuru formlarının ve diğer dokümanların bir otomasyon çerçevesinde yazılımlar tarafından işlenmesinde önemli işlerden biri imzanın bulunduğu bölgenin bulunmasıdır. İmza analizi doküman analizi ve tanıma alanında araştırmacıların büyük dikkatini çekmiş olmasına rağmen, bu çalışmalar imzanın tespiti değil doğrulanması yönünde zenginleşmektedir [15, 16, 17]. Doğrulama işleminde, kişinin imzasının daha önce alınmış imzalara göre modellenmesi gerekmektedir. Dokümanda imzanın varlığını veya pozisyonunu analiz eden çok az sayıda deneme yapılmıştır.

Taranmış dokümanlarda imzanın bulunması işinde birtakım zorluklar bulunmaktadır. İlk olarak bazı doküman tiplerinin düşük çözünürlükte olması iyileştirme işlemini zorlaştırmaktadır. İkinci olarak, her dokümanın arka planı değişken olup genellikle önceden bilinmemektedir. Ayrıca uygulamanın gereksinimleri açısından dokümanların çok kısıtlı bir zamanda işlenebilmesi gerekmektedir. Son ve belki de en önemli olarak, dokümanlarda bulunan çeşitli çizgi ve el yazılarının imzalara benzeyebilmesi veya imzalarla kesişebilmesidir.

İlk çalışmalardan birinde, insanın görsel algılamasını taklit etme yöntemine dayayan bir intuitive yaklaşım ile imza çıkarım problemi ele alınmıştır [18]. Atılan

imzalarda kıvrımların karakteristikleri için bir kriter olarak ipliksilik (filiformity) kavramını tanıtmışlardır. Temiz bir şekilde taranmış banka çeklerinde başarılı olmasına rağmen, bu yaklaşım dokümanda başka ipliksi (filiform) nesneler bulunduğunda başarısız olmaktadır.

Bir başka çalışmada, kayan bir pencere kullanılarak imzanın bulunduğu bölgeyi tahmin eden ve resim bölütünü kesen bir yöntem denenmiştir [19]. Bu aşamadan sonra, kesilen bölgenin imza olup olmadığına karar vermek için, piksel bazlı yoğunluktan türetilen lokal entropi analiz edilmiştir. Bu yaklaşım gürültüyü göz ardı etmiş ve böylece yüksek yoğunluklu bölgeleri yanlış bir biçimde imza olarak ayırmıştır.

Bazı çalışmalarda bölütlenmiş bölgeleri analiz etmek için alan, dairesellik, en boy oranı, boyut ve pozisyon gibi geometrik özellikler kullanılmıştır [20, 21]. Bu özellikler Manhattan uzaklığı gibi benzerlik metrikleri ile karşılaştırılmıştır. Ayrıca gri tonda taranmış çekler üzerinde, varsayılan imza üzerine yerleştirilen bir grid içinde varyans analizine dayanan başka bir metot önerilmiştir [22].

Diğer bir yöntemde çok ölçekli çıkıntı (multi-scale saliency) özelliği kavramı, imza karakteristiklerini tanımlamak için tanıtılmıştır [23]. Bu yaklaşım aynı zamanda imza doğrulama için de kullanılmıştır [24].

Diğer bir çalışmada üç etaptan oluşan bir prosedür sunulmuştur [25]. İlk etap, kelime tabanlı özniteliklerin çıkarımı ile imza bölütleri bulur. İkinci etap makine yazısı ile kesişen imza darbelerini ayırır. Son etap iskelet analizi ile gerçek imza darbelerini sınıflandırır. Makine öğrenimi tabanlı bir sınıflandırıcı, gradyan tabanlı öznitelikler ile besleme yapılır. SURF öznitelikleri ile de bölütleme yapılmış imza blokları sınıflandırılmıştır [26].

Bir başka yaklaşımda, imza tespitinin normal olarak bir doğrulama adımı ile takip edildiği varsayılmıştır [27]. Buna göre, elde bulunan tüm imzalardan faydalanmak için bir kanıt biriktirme stratejisi uygulanmıştır. Bu yaklaşım ile yüksek bir doğruluğa ulaşılmasına rağmen, başvuru formları sıklıkla yeni müşteriler tarafından doldurulduğundan ve sonraki adım doğrulama değil arşivleme adımı olduğundan yaklaşımları genel olarak doğru değildir.

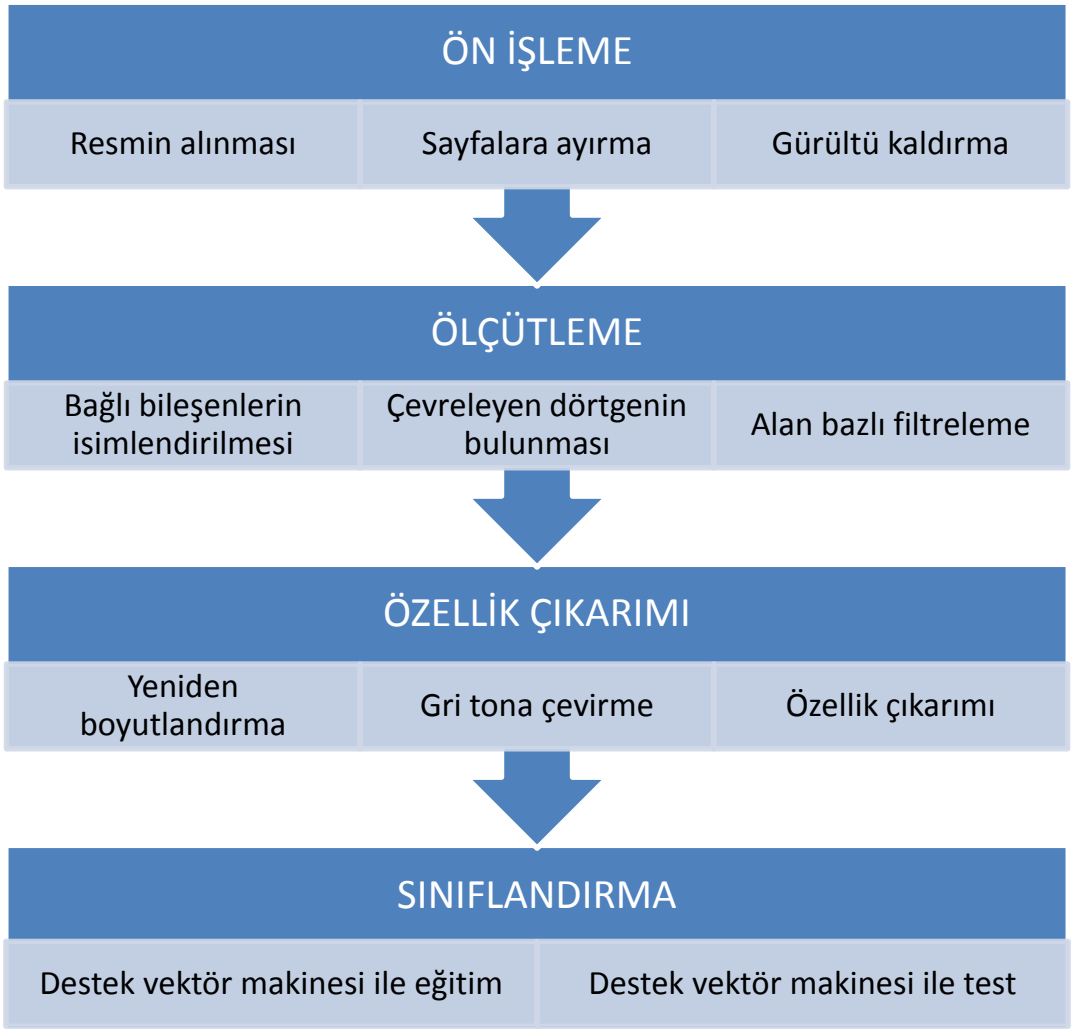
Tez kapsamında yapılan imza bölge analizi çalışmasında, çok sayfalı başvuru dokümanlarına elle atılan imzaları bulup sayfadan çıkartan bir altyapı geliştirilmiştir. Başvuru formlarının siyah/beyaz tarandığı ve müşterilerin mevcut veritabanında daha önce bir imzası bulunmadığı varsayılmıştır.

Altyapı imza olan ve olmayan örnek kümelerinden model parametrelerinin öğrenildiği bir yapı sunduğundan ayırt edicidir. Öğrenme modeli destek vektör makineleri üzerine oluşturulmuştur. Destek vektör makinelerini besleyen öznitelik kümeleri için farklı lokal ve global resim öznitelik temsil yöntemleri seçilmiştir. Gerçek veri kümeleri üzerinde yapılan denemeler, altyapının gerçek hayat uygulamalarında kullanılabilecek, makul olan iyi bir doğruluk oranına ulaşabildiğini göstermektedir. Aynı zamanda imza çıkarım probleminde kullanılan farklı resim temsil özniteliklerinin karşılaştırmalı bir analizi yapılmıştır.

3.2 Yöntem

İmza bölge analizi için kullanılan altyapı Şekil 10'da gösterilmiştir.

Süreç girdi olan resmin alınması ve daha iyi bir görünümünün elde edilmesi için bir ön işlem adımı ile başlamaktadır. İkinci adımda, imza bölge adaylarını elde etmek için bir resim bölütleme yöntemi kullanılmaktadır. Üçüncü adımda aday imza bölgeleri, resim içeriğinden oluşturulan bir takım sayısal öznitelikler kümesi ile temsil edilirler. Bu öznitelik vektörleri son aşamada destek vektör makinelerine verilirler. Önerilen yapının detayları aşağıdaki gibidir.



Şekil 10 : İmza bölge analizi altyapısı

3.2.1 Ön işlem

Ön işlem süreci, resmin alınması, girdinin çok sayfalı formatta olması durumunda sayfaların çıkarılması işlemini içerir.

Resmin daha iyi görünmesi için medyan filtresi ile gürültünün kaldırılması opsiyonel olarak gerçekleştirilmektedir. Medyan filtresinin imza ile ilgili çalışmalarda taramadan kaynaklanan gürültülerin kaldırılması için kullanıldığı görülmektedir [41]. Bununla birlikte, kullanılan bazı vektör temsillerinde, gürültü kaldırma işleminin doğruluğu azalttığı görülmüş, bu kapsamda öznitelik vektör oluşturma işlemine göre yapılması karşılaştırmalı olarak analiz edilmiştir.

Belirtilen işlemlerin dışında başka bir ön işlem adımı uygulanmayarak bundan sonraki bölütleme adımına mümkün olduğu kadar çok bilginin sağlanması hedeflenmiştir.

3.2.2 Bölütleme

Bölütleme imza bölge analizinde kritik bir adımdır. Bütün doğru pozitif bölütlerin (imza içeren bölütler) bulunması istenilen bir durumdur.

Verimli bir bölütleme işleme için, iki fazlı taramaya dayalı bağlı bileşenlerin etiketlenmesi yaklaşımı izlenmiştir [28]. Bu yaklaşım üç süreçten oluşur : (1) 4 komşulu bir maske kullanılarak her piksele bir geçici bir isim verilmesi ve eşdeğer isimlerin bulunması (2) Eşdeğer isimlerin kaydedilmesi ve her eşdeğer geçici isim için bir temsil edici isim bulunması (3) her geçici ismin temsil edici isim ile değiştirilmesi. Daha verimli bir versiyon için, resimdeki piksel ve çizgiler alışılmış olan bir bir işleme yönteminin tersine ikişer ikişer işlenir.

Bağlı bileşenlerin etiketlenmesinden sonra ilgili resim bölgeleri için kendilerini kapsayacak en ufak dikdörtgenler bulunur. Bu bölgeler daha sonra imza bölütlerini kaydetmek için kullanılacaktır.

350 pikselden daha az piksel içeren bölütler işlem sırasında kaldırılır. Bu noktada esas resimde belirlenen dikdörtgen içindeki tüm pikseller seçilmez. Seçilecek piksellerin hem ilgili dikdörtgen içinde olması hem de ilgili etikete atanmış olması gerekmektedir. Bunun sebebi, imza bulunan bölütlerin genellikle daha çok piksel içermesidir. Bu yaklaşım ile, öznitelik vektörü oluşturulacak ve sınıflandırıcıya verilecek bölüt sayısı da azaltılmış olacağından, kısıtlı sürede işlemin gerçekleştirilmesi mümkün olacaktır.

3.2.3 Özniteliklerin çıkarımı

İmza çıkarımı için ayırt edici bir yaklaşım izlendiğinden, her bölüt vektörleştirilmeli ve makine öğrenimi sınıflandırıcısı bu vektör ile beslenmelidir. Bu vektörleştirme işlemi, bölütün imza veya imza değil şeklinde temsil edilebilmesi için, içerik tabanlı bazı özniteliklerin seçilmesi ve çıkarılması ile gerçekleştirilir.

Öznitelik çıkarımından önce, ilgili öznitelik çıkarım yönteminin gereksinimlerine göre bir takım ön işlemler yapılmaktadır. Bu işlemlerden biri yeniden boyutlandırmadır. Bu durumda genellikle bölüt 126x126 piksele yeniden boyutlandırılır. Gradyanların histogramı (Histogram of Gradients – HOG) yönteminde ise, bölüt 128x128 piksele yeniden boyutlandırılır.

Yine bazı yöntemlerde gri tonla girdi üzerinde vektöre çevirme işlemi yapılabildiğinden, gelen siyah/beyaz resim gri tona çevrilmektedir. Gri tona çevirme işlemi için 2x2 boyutundaki bir medyan filtresi 5 defa uygulanmaktadır. Farklı öznitelik temsil yöntemleri değerlendirilerek imzaların diğer bağlı bileşenlerden ayrıştırılması hedeflenmiştir.

Gradyan-tabanlı öznitelikler : İlk öznitelik kümesi, lokal piksel temsillerinin gradyan vektörleri ile oluşturulmasına dayanmaktadır. Bu öznitelik kümesinin oluşturmak için, 128x128 piksele yeniden boyutlandırılmış bölüt 9x9 bloklara ayrıştırılır. Bölütteki kenarlar Robert's operatörü ile bulunur. Bulunan kenarlara ait gradyanlar 16 yönde gruplara ayrılır ve her yönde gradyanın kuvveti kadar yön bazında biriktirilir. Kuvvet, açının ark tanjantı ile bulunur. 16 yön için toplanan değerlerin histogramı her 9x9 blok için hesaplanır. Her blok 16 sayıdan oluştuğu, 9x9 blok halinde resim bölümlere ayrıldığı için bu yöntemde oluşan öznitelik vektörü 9x9x16=1296 sayıdan oluşmaktadır.

Gradyanların Histogramı (HOG – Histogram of Gradients): Gradyan tabanlı bir başka popüler yöntem HOG'tur [29]. Resim 128x128 piksel olarak yeniden boyutlandırıldıktan sonra, yöntem varsayılan parametreleri ile kullanılmıştır. Her HOG vektörü 8100 öznitelikten oluşmaktadır, gradyanlar burada 9 yönde biriktirilir. İşlem, 4 hücreden oluşan 15x15 blok için yapılmaktadır.

SIFT (Scale Invariant Feature Transform) : Üçüncü öznitelik kümesi ilgi noktalarını tanımlamaya dayanmaktadır. SIFT gri tonlu resimler için anahtar noktalar olarak adlandırılan öznitelikler üretir. Bu noktalar resmin çevrilmesi, yeniden boyutlandırılması, ve kısmen ışıklandırmadaki değişime durumlarında değişmezdir [30]. Her anahtar noktanın çevresindeki bir komşulukta lokal yapıların özet bir tanımı oluşturulur. Bu özet tanım, resim yoğunluğuna ait lokal gradyan yönlerinin toplanması ile oluşturulmaktadır. Yeterli sayıda anahtar nokta bulunması durumunda öznitelik kümesi yüksek ölçüde ayırıcıdır.

LTP (Local Ternary Patterns) : LTP resimlerdeki dokuyu modellemek için kullanılan bir metottur. Yakın zamanda Suruliandi ve Ramar tarafından LBP (Local Binary Pattern)'ye [31] bir uzantı olarak sunulmuştur [32]. LBP, tekbiçimli olarak terimleştirilen birtakım lokal ikili doku örüntülerini tanımaya dayanmaktadır. Merkezdeki piksel R yarıçaplı olan dairesel bir komşulukta bulunan P pikselleri ile karşılaştırılır. Dairesel çevrenin sınırları boyunca ikili seviyede yapılan bir karşılaştırma bir tekbiçimlilik ölçüsü bulmak için kullanılır. Bu ölçü, geçişleri ortaya koyar. Tekbiçimlilik derecesi önceden tanımlanmış bir eşikten daha az olan bir örüntü, 0 ve P aralığında değişen bir etikete atanır. tekbiçimli olmayan örüntüler ise, tek bir etikete atanır (örnek olarak P+1). LBP öznitelik temsili, ele alınan bölge üzerindeki bu tekbiçimli örüntülerin ayrık oluşma histogramının bir vektöründen oluşur. LTP, ikili bir örüntü üzerinde işleme yerine üçlü bir örüntü üzerinde işlem yapma imkanı sağlar. Geçişlerin sayısını veya örüntülerin dairesel tanımlarındaki süreksizlikleri bulmaya izin verir. Örüntünün tekbiçimliliği, ritmik bir örüntüyü takip ettiği bulunan bu geçişler ile değerlendirilir. Bu örüntülerin geniş bir bölge üzerinde oluşma frekansı LTP öznitelik temsilini oluşturur.

Global öznitelikler : Son öznitelik kümesi bölütlerin global temsili ile ilgilidir. Bu küme entropi, en-boy oranı, ve enerji özniteliklerini içerir. Verilmiş bir resim bloğu i ve piksel yoğunluğu P_i için, entropi $E_i = -P_i \log P_i$ olarak tanımlanır. Entropi, bu bölgenin içerdiği global bilginin bir ölçüsüdür. Enerji ise, bölütteki tüm piksellerin yoğunluğunun karelerinin toplamının bölüt alanına bölümünden oluşur. En-boy oranı ise bölütün eninin boyuna bölünmesi ile tanımlanan bir başka global özniteliktir.

3.2.4 Sınıflandırma

Bölütün sınıflandırılması popüler makine öğrenim metodu olan destek vektör makineleri (SVM) ile gerçekleştirilmiştir. SVM yapısal risk minimizasyonu prensibini temel alarak çalışan bir ikili sınıflandırıcıdır. Eğitim fazında SVM'nin girdisi eğitim örneklerinin önceden tanımlanmış özelliklerini temsil eden n-boyutlu öznitelik vektörleridir. SVM n-boyutlu girdi uzayını daha yüksek boyutlu bir özellik uzayını doğrusal olmayacak şekilde eşleştirir. Bu yüksek boyutlu özellik uzayında bir doğrusal sınıflandırıcı oluşturulur. Öngörme fazında, doğrusal sınıflandırıcı sorulan örnek için ayırt edici bir skor üretir. İkili sınıflandırma işleminde, skor değerinin pozitif olması test örneğinin bu sınıfa ait olduğunu gösterir. Tez çalışmasının imza bölge analizi kısmında, SVM doğrusal, polinom ve radyal çekirdekleri ve LIBSVM [14] uygulamasındaki varsayılan girdi parametreleri ile kullanılmıştır.

3.3 Sonuçlar

3.3.1 Veri kümeleri

Performans değerlendirmesi için iki veri kümesi değerlendirilmiştir.

İlk veri kümesi Tobacco-800 [23] olarak adlandırılan bir karşılaştırmalı değerlendirme (benchmark) kümesinin genişletilmiş halidir. Veri kümesi 755 bölüt içermektedir. Bunların 353 tanesi imza olup diğerleri imza değildir. Veri kümesine www.baskent.edu.tr/~hogul/signds.rar. adresinden erişilebilir.

İkinci veri kümesi bir gizlilik anlaşması ile şu anda çalışan bir sigorta şirketinden elde edilen gerçek doküman resimlerini içermektedir. Bu veri kümesinde, 2670 adet çok sayfalı doküman bulunmakta olup, toplam sayfa adedi 9943'tür. 4082 sayfada en az bir imza bulunmaktadır. 5861 sayfada ise hiç imza bulunmamaktadır.

3.3.2 Denemeler için oluşturulan kurulum

İlk veri kümesindeki iş, verilen bir bölütün imza olup olmadığını tespit etmektir. İkinci veri kümesi içinse yapılan çalışma, verilen bir dokümanın imza içerip içermediğini belirlemektir.

Temelde, sınıflandırıcının çıktısı olarak, herhangi bir sayfa en az bir imza bölütü içeriyor ise, dokümanın ilgili sayfası imza içeriyor şeklinde etiketlenmektedir. İmza bölütünün doküman içerisindeki gerçek pozisyonu ise imzanın bölgesi olarak belirtilmektedir.

Değerlendirme için ilk veri kümesinde 5-katmanlı çaprazlama (k-fold cross-validation) gerçekleştirilmiştir. İmza çıkarımı altyapısı ile ilgili pratikte yeteneği görebilmek için, altyapı ilk veri kümesindeki pozitif ve negatif örnekleri de içeren tüm bölütler ile eğitilmiş ve ikinci bağımsız veri kümesindeki dokümanlar üzerinde çalıştırılmıştır. Her iki veri kümesinde ortak doküman bulunmamaktadır.

Performans değerlendirmesi, doğruluk (doğru öngörülen etiketlerin oranı), duyarlılık (pozitif olan ve bu şekilde öngörülen örneklerin oranı), özgüllük (negatif olan ve bu şekilde öngörülen örneklerin oranı) ölçülerek gerçekleştirilmiştir. Burada

örnek, ilk veri kümesi için bir bölüte, ikinci veri kümesi için de doküman sayfasına tekabül etmektedir.

Yapılan denemeler, öznitelik temsil yöntemleri, gürültü kaldırma uygulaması, kullanılan SVM çekirdekleri gibi değişen sınıflandırma modelleri ile gerçekleştirilmiştir. Her model için, en iyi doğruluğu veren çekirdeğin sonuçları raporlanmıştır.

İkinci veri kümesinde bulunan 9943 sayfanın her modelde gözle kontrol edilmesi çok maliyetli bir işlem olacağından, bu işlemin otomatik yapılabileceği bir altyapı oluşturulmuştur. Geliştirilen bir uygulama ile, ikinci veri kümesinde bulunan dokümanların seçilip görüntülenmesi, seçilen dokümanın sayfalarına ayrıştırılması ve ayrıştırılan sayfanın görüntülenmesi sağlanmıştır. Görüntülenen sayfada imza olup olmadığı bilgisi kullanıcı tarafından belirlendikten sonra bu bilginin girilmiş ve girilen bilgiler bir veritabanında kaydedilmiştir. Bu yöntem ile, 9943 sayfanın her birisinde imza olup olmadığı bir kere kaydedilmiş, geliştirilen her yeni modele ait test sonuçlarına ait metrikler veritabanı komut dizileri (script) ile gerçekleştirilmiştir.

Yapılan bu işlem ile ilgili kullanılan teknoloji aşağıdaki şekildedir. Bilgi giriş uygulaması Microsoft .NET C# dilinde geliştirilmiş, sayfalara ayırma ve hızlı görüntüleme işlemlerinin gerçekleştirilmesi için Windows Form tipinde bir proje kullanılmıştır. Veritabanı olarak, Microsoft SQL Server kullanılmıştır. Test sonuçlarına ait metriklerin hesaplanması için Transact-SQL dilinde komut dizileri (script) kullanılmıştır.

3.3.3 Sonuçlar

Çizelge 10 imza bölge analizi işleminin ilk veri kümesi üzerindeki sonuçlarını göstermektedir.

Çizelge 10 : İmza tespit işleminin ilk veri kümesi üzerindeki sonuçları

Öznitellik	Gürültü kaldırma uygulandı	SVM çekirdeği	Duyarlılık	Özgüllük	Doğruluk
Gradyan	Evet	doğrusal	87,0	92,8	90,1
	Hayır	doğrusal	90,1	93,0	91,7
HOG	Evet	doğrusal	39,9	94,0	68,7
	hayır	doğrusal	35,1	93,8	66,4
SIFT	evet	doğrusal	70,5	76,9	73,9
	hayır	doğrusal	77,9	75,9	76,8
LTP	evet	polinom	93,8	90,0	91,8
	hayır	polinom	92,9	91,0	91,9
Global	evet	polinom	58,9	40,8	49,3
	hayır	polinom	71,7	66,9	69,1
Gradyan + Global	hayır	doğrusal	94,9	95,0	95,0
LTP + Global	hayır	doğrusal	92,9	93,0	92,9
Gradyan + LTP + Global	hayır	polinom	94,1	91,8	92,8

Çizelge, tek öznitelik temsillerini ve kombinasyonlarına ait sonuçları içermektedir. SIFT özniteliklerinin imza tespiti ile ilgili istenilen noktada olmadığı açıktır. Sadece SIFT öznitelikleri dikkate alındığında, tüm performans metrikleri diğerlerine göre çok aşağıda kalmaktadır.

HOG özelliği, en yüksek özgüllüğü çok düşük bir duyarlılık ve düşük doğrulukla sağlamaktadır.

Gradyan bazlı ve LTP öznitelikleri en yüksek doğruluk seviyelerine ulaşırken, duyarlılık ve özgüllük arasında mantıklı bir denge oluşturmuşlardır. Burada LTP yönteminin biraz daha iyi sonuç verdiği görülmektedir.

Bununla birlikte, salt LTP kullanılarak elde edilen performans, diğer özniteliklerin eklenmesi ile oluşturulan entegre bir öznitelik kümesi ile geliştirilememektedir.

Global öznitelikler tek başına kullanıldığı zaman yararsız gözükmekle birlikte, gradyan-tabanlı öznitelikler ile birlikte kullanıldığında toplam doğruluğu geliştirebilmektedirler. Bu kombinasyon tüm metrikler için en yüksek performansa ulaşmaktadır.

Enteresan bir sonuç olarak gürültü kaldırma işlemi, imza bölge analizi performansı için yararlı değil hatta zararlı olabilmektedir. Bu durum yüksek olasılıkla, imzanın içindeki bağlı çizgilerin ve kenarlarda bulunan farklı eğimlerin gürültü kaldırma işlemi ile kısmen kaldırılması ve imza ayrıştırma işleminde tanımlayıcı olan bu bilgilerin bu sebepten kullanılamaması olarak açıklanabilir. Bu durum bilindiğinden, entegre edilmiş öznitelik kümelerinde ve ikinci veri kümesinde gürültü kaldırma işlemi uygulanmamıştır.

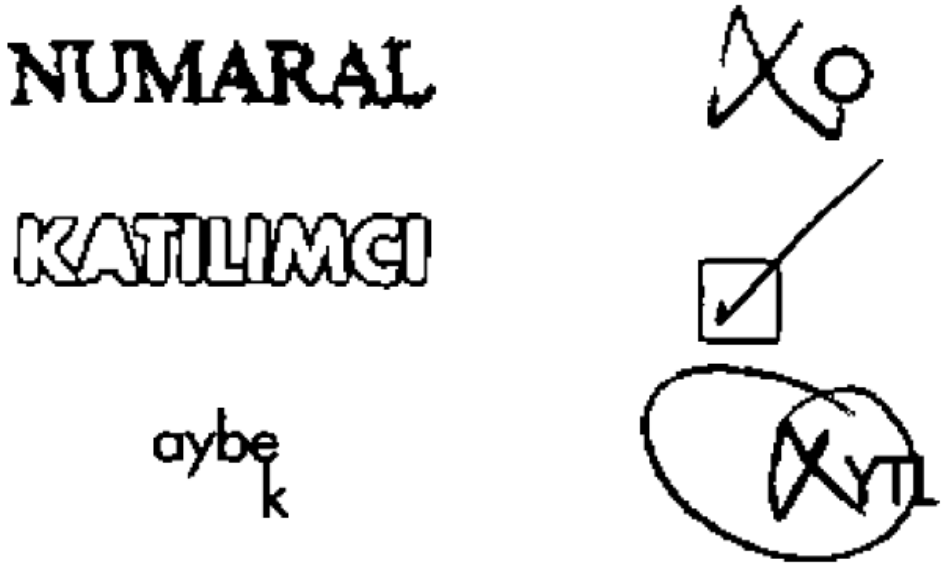
Çizelge 11'de sayfanın bir imza içerip içermediğinin belirlenmesi işlemi ile ilgili denemelerin sonuçları gösterilmiştir. Bu deneme, tüm doküman sayfalarını içeren ikinci veri kümesi ile yapılmıştır. Sınıflandırıcı ise, birinci veri kümesinde bulunan imza örnekleri ile eğitilmiştir. Sonuçlar yine gradyan-bazlı öznitelik kümelerinin global öznitelikler ile birleştirildiğinde, taranmış dokümanlar üzerinde imza bulma işlemi ile ilgili olarak en güvenilir yöntemi oluşturduğu göstermektedir.

Çizelge 11: İmza tespit işleminin ikinci veri kümesi üzerindeki sonuçları

Öznitelikler	Duyarlılık	Özgüllük	Doğruluk
Gradyan	93,8	47,7	66,6
LTP	97,6	29,7	57,5
SIFT	98,5	16,3	51,2
Global Öznitelikler	99,8	14,3	49,3
Gradyan+Global	95,0	54,3	71,0
Gradyan+Global + LTP	94,9	44,8	65,4
LTP + Global	97,8	29,1	57,3

İmza bölge analizindeki bazı problemler, sonuçların elle incelenmesi sonucu ortaya çıkmaktadır. Yanlış pozitifler için iki temel sebep olduğu görünmektedir : (1)

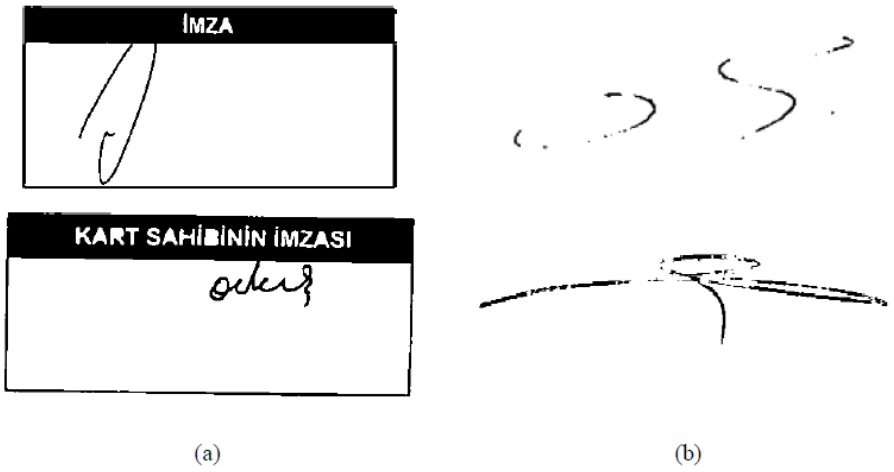
Birbiri ile bağlantılı fontlardan oluşan makine yazısı metinler (2) el yazısı harfler veya onay işaretleri vb. işaretler. Bu sebepler ile ilgili şekiller Şekil 11'de gösterilmiştir.



(a) (b)

Şekil 11 : Sık rastlanan yanlış-pozitifler (a) makine çıktısı metinler (b) El yazısı harf veya işaretler

Yanlış negatif ise, genelde imzaların dokümandaki diğer metin veya şekiller ile kesişmesi sonucu ortaya çıkmaktadır. Bu durum özellikle ilgili imzanın kesiştiği şekle göre küçük kalması durumunda ortaya çıkmaktadır. Bu durum Şekil 12'de gösterilmektedir. Aynı zamanda burada bir başka problem, tarama veya doküman kalitesinden ötürü dokümandaki imzanın süreksiz veya kesikler halinde oluşmasıdır. Özellikle faks ortamında oluşan dokümanlarda bu problemin çıkabildiği görülmektedir. Taranan dokümanlarda bu problem ise, tarama çözünürlüğünün düşük olması, pellur vb. ince kağıtlardaki görüntü kalitesinin daha kötü olabilmesi, tarayıcının kağıdı tarama sırasında alırken oluşabilecek kaymalardan ötürü doküman kalitesinin bozulması sayılabilir.



Şekil 12 : Sık rastlanan yanlış-negatifler (a) İmzanın diğer bilgi ile kesişmesi (b) Kesikli imzalar

4. TARTIŞMA VE GELECEK ÇALIŞMALAR

Tez çalışmasında, resim tabanlı dokümanların kategorize edilmesi ve doküman üzerinde imzanın tespit edilmesi ile ilgili bir altyapı sunulmuştur.

Resim tabanlı dokümanların kategorize edilmesi için sunulan altyapı, metin özniteliklerini, resim özniteliklerini ve her iki öznitelik kümesini aynı zamanda kullanarak verilen bir dokümanı bilinen sınıflardan birisine kategorize etmektedir. Modeller Türkiye'deki bankacılık iş süreçlerinde sıklıkla kullanılan dokümanlar için özelleştirilmiştir. Bu kapsamda, Türkçe dil desteği bulunan OCR motorları kullanılmıştır. Elde edilen sonuçlar önerilen altyapının gerçek uygulamalarda kullanılabilecek düzeyde olduğunu göstermektedir.

Doküman kategorizasyonu ile ilgili gelecek çalışmalar içerisinde, resim öznitelikleri ve metin özniteliklerinin kullanımının genişletilmesi düşünülmektedir. Özellikle resim öznitelikleri ile kategorizasyon sürecinde, kelime şekil kodlama (word shape coding) gibi yöntemler kullanılarak, aslında OCR işlemi uygulanmamış bir dokümanda belirli kelimelerin varlığının yüksek hızlı kontrol edilebilmesi değerlendirilmektedir. Yapılan ön incelemede, literatürde kullanılan çeşitli kelime şekil kodlama yöntemlerinin Türkçe karakter kümesinde bulunan Türkçe harfleri içermediği görüldüğünden, bu yaklaşımda, kodlama yönteminde geliştirmeler yapılması gerektiği öngörülmektedir. Resim tabanlı yaklaşıma bu özelliğin eklenmesi, sadece şekilsel olarak birbirine benzeyen dokümanların değil, şekilsel olarak birbirine benzeyen ancak içerik farkı ile ayrıştırılabilen bazı dokümanların da kategorize edilebilmesini sağlayacaktır.

Dokümanın resim tabanlı yöntemlere kategorize edilmesinde logo, amblem veya çeşitli şekillerin tanınması ile altyapının zenginleştirilmesi planlanmaktadır. Doküman üzerinde belirli bir şeklin belirli bir boyutta bulunması durumunda, dokümanın kategorisi hakkında bilgi edinilebilecek şekilde dokümanlar bulunduğundan, yapılacak bu çalışmanın da başarı oranını arttıracağı değerlendirilmektedir.

İmza bölge analizi ile ilgili olarak, resim tabanlı dokümanlarda imza tespitini gerçekleştiren bir altyapı sunulmuştur. Altyapı, sağlam, güvenilir ve yüksek

performansla çalışan bir bölütleme adımı içermektedir. Çeşitli resim özelliği kullanılarak imza olan ve olmayan resimler temsil edilmeye çalışılmıştır. Denemelere dayanan sonuçlara bakıldığında, gradyan ve LTP tabanlı özniteliklerin tek başlarına kullanıldıklarında imza içeren bölütleri daha yararlı olduğu görülmüştür. Bazı durumlarda, öznitelik temsil yöntemlerinin birleştirilmesi tahminlerin güvenirliğini arttırmaktadır. Bu kapsamda, en boy oranı, enerji ve entropi gibi global öznitelikler, aday bölütlerin temsil edilmesinde tamamlayıcı öznitelik olabilmektedirler. Gerçek hayatta kullanılan sigorta dokümanlarında yapılan denemelerle ilgili sonuçlar, geliştirilen altyapının gerçek uygulamalarda kullanılmasını motive edicidir. Altyapı, ek lokal resim öznitelikleri ile genişletilebilir durumdadır.

Gelecek çalışmalarda, imza bölge analizi kapsamında geliştirilen altyapının kurallar ile genişletilebilmesi planlanmaktadır. Özellikle çok sayfalı dokümanlarda imzanın hangi sayfalarda bulunabileceğinin kurallarla belirtilebilmesi başarıyı arttırabilecektir. Ek olarak, sayfanın tamamı değil, belirli bölgelerinin kontrol edilmesi başarıyı arttırabilecek bir başka yaklaşımdır. Doküman kategorizasyon süreci çalıştırılarak doküman tipinin belirlenmesi, imza tespiti yapılmak istenen doküman hakkında daha çok bilgi sağlayacağından, imza tespit süreci ile birlikte çalıştırılabilecektir.

İmza bölge analizi ile yapılacak gelecekteki bir diğer çalışma, bağlı bileşenlerin etiketlenmesi yöntemi ile elde edilen bölütlerin büyük olması durumunda, ikinci bir bölütleme yöntemi kullanılarak imzanın daha küçük bölgelerde aranması olacaktır. Özellikle imzanın resimdeki başka çizgilere değerek büyük bölütlerin içinde kalması durumunda oluşan imzanın tespit edilememesi durumu böylece belirli bir oranda çözülerek başarının arttırılacağı düşünülmektedir.

KAYNAKLAR LİSTESİ

[1] Greg Goth: A Structure for Unstructured Data Search, IEEE Distributed Systems Online, vol. 8, no. 1, 2007, art. no. 0701-o1003 (2007)

[2] Keith Kmetz : Scanning Brings Cost and Efficiency Benefits to a Wide Range of Common, Everyday Business Processes, IDC Report, (2006)

[3] Türk Standartları Enstitüsü : TS 13298 Elektronik Belge Yönetimi, Türk Standardı, Haziran 2009 (2009)

[4] Lam, S.: An adaptive approach to document classification and understanding. In: Proceedings of International Association for Pattern Recognition Workshop on Document Analysis Systems, Kaiserslautern, Germany, October 1994, pp. 231–251 (1994)

[5] Baumann, S., Ali, M., Dengel, A., Jäger, T., Malburg, M., Weigel, A.,Wenzel, C.: Message extraction from printed documents– a complete solution. In: Proceedings of the 4th International Conference on Document Analysis and Recognition, Ulm, Germany, 18–20 August 1997, pp. 1055–1059 (1997)

[6] Hu, J., Kashi, R., Wilfong, G.: Document classification using layout analysis. In: Proceedings of the 1st International Workshop on Document Analysis and Understanding for Document Databases, Florence, Italy, September 1999, pp. 556–560 (1999)

[7] Wnek, J.: Learning to identify hundreds of flex-form documents. In: Proceedings of Document Recognition and Retrieval VI (IS&T/SPIE electronic imaging), San Jose, CA, 27 January 1999, SPIE Proceedings Series 3651, 173–182 (1999)

[8] Cesarini, F., Lastri, M., Marinai, S., Soda,G.: Encoding of modified X–Y trees for document classification. In: Proceedings of the 6th International Conference on Document Analysis and Recognition, Seattle, USA, 10–13 September 2001, pp. 1131–1136 (2001)

[9] Spitz, A.L., Maghbouleh,A.: Text categorization using character shape codes. In: Proceedings of Document Recognition and Retrieval VII (IS&T/SPIE electronic imaging), San Jose, California, 23–28 January 2000, SPIE Proceedings Series 3967, 174–181 (2000)

[10] Baldi, S.,Marinai,S., Soda,G.: Using tree-grammars for training set expansion in page classification. In: Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, 3–6 August 2003, pp. 829–833 (2003)

[11] Diligenti, M., Frasconi,P., Gori, M.: Hidden Tree Markov Models for document image classification. IEEE Trans. Pattern Anal. Mach. Intell. 25(4), 519–523 (2003)

[12] Ogata, H., Watanabe, S., Imaizumi, A., Yasue, T., Furukawa, N., Sako, H., Fujisawa, H.: Form type identification for banking applications and its implementation issues. In: Proceedings of Document Recognition and Retrieval X (IS&T/SPIE electronic imaging), Santa Clara, California, 20–24 January 2003, SPIE Proceedings Series 5010, 208–218 (2003)

[13] Kumar J, Pillai J, Doermann D, Document Image Classification and Labeling using Multiple Instance Learning, Intl. Conf. on Document Analysis and Recognition (2011).

[14] C.Chang, C.Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, vol.2.27, s.1-27, 2011.

[15] D. Impedovo and G. Pirlo. Automatic Signature Verification: The State of the Art. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, 38:5. 2008.

[16] R. Plamondon, S.N. Srihari. On-line and off-line handwriting recognition: a comprehensive survey. IEEE Trans. Pattern Anal.Mach. Intell., 22(1), 63–84. 2000.

[17] W.K.H. Weiping, Y. Xiufen. A survey of off-line signature verification. International Conference on Intellingence Mechatronics and Automation. 2004.

[18] S. Djeziri, F. Nouboud, R. Plamondon. Extraction of signatures from check background based on a filiformity criterion. IEEE Trans. Image Process. 7(10), 1425–1438, 1998.

[19] V.K. Madasu, B.C. Lovell. Automatic segmentation and recognition of bank cheque fields. Digit. Image Comput. Tech. Appl., 80(1): 33–40. 2005.

[20] V.K. Madasu, M.H.M. Yusof, M. Hanmandlu, K. Kubik. Automatic extraction of signatures from bank cheques and other documents, DICTA'03. 2003.

[21] A. Chalechale, G. Naghdy, P. Premaratne, A. Mertins. Document Image Analysis and Verification Using Cursive Signature. IEEE International Conference on Multimedia and Expo. 2004.

[22] R. Jayadevan et al. Variance based extraction and hidden Markov model based verification of signatures present on bank cheques. International Conference on Computational Intelligence and Multimedia Applications. 2007.

[23] G. Zhu, Y. Zheng, D. Doermann, S. Jeager. Multi-scale Structural Saliency for Signature Detection. IEEE Conference on Computer Vision and Pattern Recognition. 2007

[24] G. Zhu, Y. Zheng, D.Doermann, S. Jaeger. Signature detection and matching for document image retrieval. IEEE Trans. Pattern Anal. Mach. Intell., 31, 2015–2031. 2009.

[25] R. Mandal, P.P. Roy, U.Pal. Signature Segmentation from Machine Printed Documents using Conditional Random Field. International Conference on Document Analysis and Recognition. 2011.

[26] S. Ahmed, M.I. Malik, M. Liwicki, A. Dengel. Signature Segmentation from Document Images. International Conference on Frontiers in Handwriting Recognition. 2012.

[27] J.L. Esteban, J.F. Vélez, Á. Sánchez. Off-line handwritten signature detection by analysis of evidence accumulation. IJDAR, 15:359–368. 2012.

[28] L. He, Y. Chao, K. Suzuki. A New Two-Scan Algorithm for Labeling Connected Components in Binary Images. Proceedings of the World Congress on Engineering. 2012.

[29] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2005.

[30] D.G. Lowe. (1999). Object recognition from local scale-invariant features. 7th International Conference on Computer Vision. 1999.

[31] T. Ojala, M. Pietikäinen, D. Harwood. A Comparative Study of Texture Measures with Classification Based on Feature Distributions. Pattern Recognition, 29:51-59. 1996.

[32] A. Suruliandi, K.Ramar. Local Texture Patterns –A Univariate Texture Model for Classification of Images. 16th International Conference on Advanced Computing and Communications. 2008.

[33] D.Boswell, Introduction to Support Vector Machines, 2002.

[34] V.Vapnik, C.Cortes, Support vector networks, Machine Learning, vol. 20, s. 273-297, 1995

[35] D. Huson, SVMs and Kernel Functions, Algorithms in Bioinformatics II SoSe‟07 ZBIT, s.265, 2007

[36] R.Kong, B.Zhang, Autocorrelation Kernel Functions for Support Vector Machines, Third International Conference on Natural Computation, 2007.

[37] T.Howley, M.G.Madden, The Genetic Kernel Support Vector Machine: 38 Description and Evaluation, Artificial Intelligence Review, vol.24, no.3-4, s.379-395, 2005.

[38] C.Hsu, C.Chang, C.Lin , A Practical Guide to Support Vector Classification, 2003

[39] Y.Abu-Mostafa, Lecture14 - Support Vector Machines, 2012, http://www.youtube.com/watch?v=eHsErlPJWUU, Erişim Tarihi : 10.08.2014

[40] Y.Abu-Mostafa, Lecture 15- Kernel Methods, 2012, http://www.youtube.com/watch?v=XUj5JbQihlU&feature=relmfu%20kernel%20methods, Erişim Tarihi : 10.08.2014

[41] A. C. Verma, D. Saha, H. Saikia. Forgery Detection in Offline Handwritten Signature Using Global and Geometric Features, International Journal of Computer and Electronics Research, Volume 2 Issue 2, April 2013

[42] Tesseract-ocr (Version 3.02) [Software]. (2013). Google. https://code.google.com/p/tesseract-ocr/

[43] Abby Fine Reader Engine (Version 10) [Software]. (2012). Abby.

[44] Magnetic ink character recognition, http://en.wikipedia.org/wiki/Magnetic_ink_character_recognition, Erişim Tarihi : 10.08.2014