Türkçe Kural Tabanlı Resmi Doküman Tipi Tespiti
Turkish Rule-Based Official Document Type Detection

Bekir BAKAR*, Filiz AKSOY†, Apdullah YAYIK‡,
Sevcan İÇÖZ§ ve Vedat AYBAR¶
Mobildev
İstanbul, Türkiye
{bekir.bakar*, filiz.aksoy†, apdullah.yayik‡, sevcan.icoz§, vedat.aybar¶}@mobildev.com

Tunga GÜNGÖR
Bilgisayar Mühendisliği Bölümü
Boğaziçi Üniversitesi
İstanbul, Türkiye
gungort@boun.edu.tr

***Özetçe* —Bu çalışma, kurumların 2016 yılında Türkiye'de yürürlüğe giren 6698 sayılı Kişisel Verilerin Korunması Kanununa (KVKK) uyum sağlamalarına yardımcı olmak amacıyla, resmi dokümanlardaki kişisel bilgileri ve aralarındaki ilişkileri tespit etmek için geliştirilen DATAMIN isimli ürünün ilk aşamasını kapsamaktadır. Dokümanlarda bulunan alan isimlerinin ayırt edici etki değerleri belirlenerek, esnek olarak tasarlanmış düzenli ifadeler ve en az düzeltme mesafesi ile eşleşme kontrolüne dayalı, kural tabanlı resmi doküman tipi tespitine yönelik metot geliştirilmiştir. Önerilen metodun kaliteli optik karakter tanıma işleminin mümkün olduğu durumlarda son derece etkili olduğu ve doğru modelleme yapabildiği tespit edilmiştir.**

Anahtar Kelimeler—KVKK, DATAMIN, doküman tipi tespiti, düzenli ifadeler, en az düzeltme mesafesi.

***Abstract*—This study is the first stage of industrial application that will be used in the product named DATAMIN, which is being developed to help companies adapt Personal Data Protection Law (DPL) No. 6698 came into force in 2016 in Turkey, by extracting and relationing personal information in official documents. Rule-based official document type detection method based on matching control with flexible regular expressions and minimum edit distance was developed by determining the distinctive effect values of the field names in the documents. It was found that proposed method was highly effective and able to make accurate modeling when optical character recognition with high-quality was avaliable.**

Keywords—DPL, DATAMIN, document type detection, regular expressions, minimum edit distance.

I. GİRİŞ

Kurum ve kuruluşlar çeşitli işlemleri yerine getirebilmek için, temas kurdukları şahıslara ait kişisel verileri kayıt altına almakta ve işlemektedir. Bu süreçler, başta özel hayatın gizliliği olmak üzere kişilerin temel hak ve özgürlüklerini korumak için kişisel verileri işleyen gerçek ve tüzel kişilerin yükümlülüklerini belirlemek amacıyla Avrupa Birliği ülkelerinde Genel Veri Koruma Kanunu −General Data Protection Regulation, GDPR [1], Türkiye Cumhuriyetinde ise 6698 sayılı Kişisel Verilerin Korunması Kanunu (KVKK) ile [2] düzenlenmiştir. Bahsi geçen kanunların getirdiği hukuki ve ekonomik yaptırımlar, kurum ve kuruluşları kişisel bilgileri düzenleme ve kontrol altına almak için çözüm arayışına yöneltmiştir. Bu çalışmanın amacı, doğal dil işleme yöntemleri ile KVKK kapsamında kişisel veri barındıran resmi dokümanların sınıflandırılmasıdır. Kişisel veri içeren doküman tipinin belirlenmesi için kural tabanlı olarak geliştirilen sistem, her bir doküman tipine özgün öznitelikler esas alınarak tasarlanmış düzenli ifadeler ve en az düzeltme mesafesi ile içerik tabanlı eşleşme kontrolüne dayanmaktadır. İçerik tabanlı sınıflandırılması amaçlanan dokümanların çoğunluğu e-devlet üzerinden temin edilebilen ve kişisel bilgi içeren 11 adet doküman tipidir. Bu dokümanlar aşağıda belirtilmiştir.

- Eski / Yeni Kimlik Belgesi (D-1 / D-2)
- Eski / Yeni Sürücü Belgesi (D-3 / D-4)
- Adli Sicil Belgesi (D-5)
- Öğrenci Belgesi (D-6)
- Nüfus Kayıt Belgesi (D-7)
- Araç Ruhsatı (D-8)
- Mezun Belgesi (D-9)
- Yerleşim Yeri ve Adres Bilgileri Belgesi (D-10)
- Askerlik Durum Belgesi (D-11)

burada D, doküman numarasını temsil etmektedir.

Bölüm II'de mevcut yöntemler ve endüstriyel çözümler ele alınmıştır. Bölüm III'de kullanılan metotlar ve karar verme süreci, Bölüm IV'de ise deneysel sonuçlar ve değerlendirmeler bulunmaktadır.

II. GEÇMİŞ ÇALIŞMALAR

Geçmiş yıllarda birçok doküman sınıflandırma yöntemi önerilmiştir. Bunlardan en çok karşılaşılan resim türündeki dokümanların yapısal yerleşim benzerliğini kullanan çalışmalardır. Bu yöntem kullanılarak sayısal resim makina öğrenmesi tabanlı yaklaşımlar ile sınıflandırılmaktadır [3]–[5]. Ayrıca resim türündeki dokümanlar için optik karakter tanımaya (OKT) bağımlı olan içerik metni tabanlı benzerlik kullanımına dayalı çalışmalar da mevcuttur. Bu yöntemde ise terim frekansı−ters doküman frekansı (TF−TDF) metodu ile kural tabanlı [6] yaklaşımlar kullanılmaktadır. Ayrıca makina öğrenmesi yöntemlerinden k-en yakın komşuluk [3], destek vektör makinaları [5], naïve bayes [7] ile de sınıflandırmalar yapılmaktadır. Han ve arkadaşları [6] tarafından önerilen metot, verilen bir dokümanın her bir kelimesinin TF−TDF değeri ile oluşturulan

978-1-7281-7206-4/20/$31.00 ©2020 IEEE

ve o dokümanı temsil eden doküman vektörünün daha önceden belirlenen doküman tipleri için oluşturulmuş olan doküman vektörleri ile olan kosinüs benzerliği bilgisine dayalıdır. Bu kural tabanlı çalışmanın deneysel sonuçlarına bakıldığında makina öğrenmesi yaklaşımlarından daha üstün olduğu görülmektedir. Reklam, mektup, özgeçmiş gibi 16 adet sınıfa ait toplam 400.000 adet resim içeren RVFL-CDIP [8] veri seti ile 2017 yılında Chris ve arkadaşları [9] evrişimli yapay sinir ağı modeli ile başarılı sınıflandırma gerçekleştirmişlerdir. Türkçe kişisel bilgi içerikli resmi doküman sınıflandırma problemi için mevcut benzer bir veri setinin bulunmadığı, bu sebeple benzer çalışmaların yapılamadığı gözlemlenmiştir. Endüstride karşılaşılan IBM firmasının Stored IQSuite, Amazon firmasının Macie, Sipiron firmasının ve Titus firmasının Illuminate ürünlerinin, Türkçe resmi doküman sınıflandırmaya çözüm getirmediği görülmüştür.

III. Materyal ve Metotlar

Kamera veya tarayıcı aracılığıyla digital ortama aktarılan dokümanlarda genişleme, doğrusal normalizasyon ve Otsu adaptif eşikleme [10] metotları ile iyileştirmeler yapılmıştır. Ardından tesseract[1] OKT kütüphanesi ile bilgi çıkarımı yapılmıştır. İlerleyen alt bölümlerde kullanılan metotlar, etki faktörü tanımlaması ve karar verme süreçleri anlatılmıştır.

A. Düzenli İfadeler ve En Az Düzeltme Mesafesi

Düzenli ifadeler esas olarak karakter dizisi tespiti için cebirsel gösterimlerden oluşmaktadır. Büyük veri ambarında örüntüsü belirli olan kelime yapıları tespiti ve geri dönüşümü için kullanılmaktadır [11]. En az düzeltme mesafesi ise iki farklı kelimenin biribirlerine dönüştürülebilmesi için ihtiyaç duyulan en az düzeltme işlemi (ekleme, silme ve değiştirme gibi) sayısı olarak tanımlanır [12].

B. Etki Değeri Tanımlaması

Her bir doküman için Şekil 1'de örnek olarak gösterilen alan isimleri öznitelik olarak kullanılmıştır. Her alan isminin her doküman tipi için karar verme işlemini etkileyen farklı bir belirleyici etki değeri vardır. Eğer "memleket" alan ismi sadece bir adet dokümanda mevcut ise etki değeri, birden fazla dokümanda mevcut olan "soyadı" alan isminin etki değerinden yüksek olmalıdır. Ayrıca, bu değer toplam 10 adet alan ismi bulunan mezuniyet belgesi için 8 adet alan ismi bulunan kimlik belgesinden daha düşük olmalıdır. Alan isimlerinin etki değerinin ne seviyede olduğu ancak ilgili uzayının bir arada değerlendirilmesi ile belirlenebilir. Bu durumda aynı alan isimlerinin farklı doküman türlerinde mevcut olabilme durumu (TDF) ve dokümanların toplam alan sayılarının göz önünde bulundurulması gereklidir.

Belirtilen gereklilikler ve kısıtlar değerlendirilerek, tüm uzayda bulunan eşsiz alan isimlerine ters frekans değeri ataması Algoritma 1'de gösterildiği gibi yapılmıştır. Burada, $\mathbf{A}$ benzersiz alan isimlerini, $\mathbf{D}$ doküman tiplerine ait alan isimlerini ve $\mathbf{F}$ ise benzersiz alan isimlerine ait ters frekans değerlerini temsil etmektedir. $\mathbf{D}$'nin her elemanı bir adet doküman tipine ait alan isimlerini içeren dizidir. Örneğin Şekil 1-b'deki TC kimlik kartı doküman tipi d_1 dizisi ile temsil

[1]https://github.com/tesseract-ocr/tesseract

(a)

(b)

Şekil 1: Alan isimleri yeşil dörtgen ile gösterilen 2 adet örnek doküman; (a) mezuniyet belgesi, (b) TC kimlik kartı

Algoritma 1: Ters Frekans Değerleri Tespiti

Girdi: $\mathbf{A} = [a_1, a_2, \ldots, a_n], \mathbf{D} = [d_1, d_2, \ldots, d_m]$
Çıktı: $\mathbf{F} = [f_1, f_2, \ldots, f_n]_{ilk\_atama=0}$
for *each* a_i **do**
  for *each* d_j **do**
    If a_i *in* d_j
      $f_i \leftarrow f_i + 1$
    end
  end
end

edilmektedir. Bu dizinin ilk 3 elemanı ise şöyledir: d_{11} = TC Kimlik No/Identity No, d_{12} = Soyadı/Surname, d_{13} = Ad/Given Name. Sadece Şekil 1'de bulunan 2 adet doküman tipi birlikte değerlendirildiğinde ters frekans değeri "TC Kimlik No" alan ismi için 2, "Diploma No" ve "Geçerlilik Tarihi" alan isimleri için ise 1 olmaktadır.

Algoritma 1'de belirlenen her alana ait $\mathbf{F}$ ters frekans değerleri kullanılarak Algoritma 2'de herbir doküman tipinde bulunan alan isimlerinin etki değerleri ($\mathbf{E}$) hesaplanmıştır;

$$e_i = \frac{uzunluk(\mathbf{D})}{\mathbf{F}(d_{ij}) \times uzunluk(\mathbf{d_i})} \tag{1}$$

burada $uzunluk(\mathbf{D})$ doküman tipi sayısını, $\mathbf{F}[d_{ij}]$ ters frekans değerini ve $uzunluk(d_i)$ değeri doküman tipinin alan ismi sayısını belirtmektedir. Klasik TF−TDF de bulunan logaritma işleminin kullanılmamasının sebebi etki değerlerinin 0 veya negatif olmasını engellemektir. Şekil 1-b'deki TC kimlik kartı dokümanına ait tüm alan isimlerinin ağırlıkları e_1 dizisi ile temsil edilmektedir. Sadece Şekil 1'deki doküman tipleri bir arada değerlendirildiğinde Algoritma 1 ve 2 kullanılarak elde edilen örnek etki alanları Tablo I'de gösterilmiştir.

Algoritma 2: Belirleyici Etki Değeri Tespiti

Girdi: $\mathbf{D} = [d_1, d_2, \ldots, d_m], \mathbf{F} = [f_1, f_2, \ldots, f_n]$
Çıktı: $\mathbf{E}$
for *each* d_i **do**
  $e_i \leftarrow [e_1, e_2, \ldots, e_{h=uzunluk(d_i)}]_{ilk\_atama=0}$
  for *each* d_{ij} **do**
    $e_{ij} \leftarrow uzunluk(\mathbf{D})/(\mathbf{F}[d_{ij}] \times uzunluk(d_i))$
  end
end

TABLO I: Örnek Etki Değerleri

Doküman Tipi	Alan Sayısı	Alan Adı	Ters Frekans	Etki Değeri
D-2	8	TC Kimlik No	2	$\frac{2}{2\times8} = 0.125$
		Geçerlillik Tarihi	1	$\frac{2}{1\times8} = 0.250$
D-10	10	TC Kimlik No	2	$\frac{2}{2\times10} = 0.100$
		Diploma No	1	$\frac{2}{1\times10} = 0.200$

C. Benzeşme Oranı Tespiti ve Karar Verme İşlemi

Benzeşme oranı belirlenmesinde, alan isimlerinden oluşturulan düzenli ifadelerin en az düzeltme mesafesi kriterlerine göre eşleşmesi ve eşleşmemesi durumu göz önünde bulundurulmuştur. Alan isimleri için düzenli ifadeler, OKT esnasında meydana gelebilecek Türkçe karakter bozuklukları göz önünde bulundurularak esnek olarak tasarlanmıştır. Örneğin Şekil 1-b'de kimlik kartında "Önceki Soyadı" alan ismi bu işlem sonrasına "[OÖ]nceki ?Soyad[ıi]" düzenli ifade örüntüsü olmaktadır. Örüntülerin tüm karakterlerinde büyük ve küçük harf olabilmesi sağlanarak eşleşmelerin kapsamı artırılmıştır. Ayrıca, oluşturulan her bir düzenli ifade örüntü öncesi ve sonrasında kelime sınırı (boşluk veya bir sonraki satıra geçiş) kısıtı eklenerek istenmeyen eşleşmeler önlenmiştir. En az düzeltme mesefesinde maliyetler her bir işlem (silme, ekleme ve değiştirme) için 1 olarak belirlenmiştir.

En az düzeltme mesafesi karakter sayısı az olan alan isimleri için yüksek seviyede belirlenir ise istenmeyen birçok eşleşmeyi hatalı olarak yapacağından sistemi yanıltacaktır. Örneğin "soyadı" alan ismi için 3 olarak belirlenirse "sanatı" kelimesi ile hatalı eşleşme meydana gelecektir (3 değiştirme). Ayrıca karakter sayısı çok fazla olan alan isimleri için çok düşük seviyede belirlenir ise sistemin esnekliği azaltılmış olur ve gerçek eşleşmeler istenmeden ihmal edilmiş olur. Örneğin "Geçerlilik Tarihi" alan ismi için 1 olarak belirlenirse "Geçerlilih Tariki" kelimeleri ile hatalı eşleşmeme meydana gelir. Bu sebeple düzeltme mesafesi deneysel incelemeler neticesinde aşağıdaki gibi belirlenmiştir,

$$\mathbf{ED}(a_i) = \begin{cases} 0 & a_i\text{'nin karakter sayısı} \leq 5 \\ 1 & 5 < a_i\text{'nin karakter sayısı} \leq 15 \\ 2 & 15 < a_i\text{'nin karakter sayısı} \end{cases} \quad (2)$$

burada, $\mathbf{ED}$ en az düzeltme mesafesini ifade etmektedir. Düzenli ifade eşleşmesi ile en az düzeltme mesafesi aynı sonlu durum otomata içerisinde birlikte çalışarak oldukça esnek ama aşırılıktan uzak bir biçimde doğru eşleşme geri dönüşümü yapılabilecek şekilde tasarlanmıştır. Doküman tipi, ilk olarak Algoritma 3 ile benzeşme oranı tespiti ve sonrasında karar verme işleminden sonra belirlenmiş olmaktadır. Algoritma 3'de tipi öğrenilmek istenilen dokümanın belirlenen doküman tiplerine ne kadar benzeştiği ölçülmektedir. Burada, $\mathbf{Dok}$ tipi belirlenmek istenilen bir dokümanı, $\mathbf{E}$ Algoritma 2'de hesaplanan etki değerlerini, $\mathbf{ED}$ en az düzeltme mesafesini ve $\mathbf{B}$ tipi belirlenmek istenilen dokümanın benzeşme oranlarını ve $\mathbf{T}$ ise her doküman tipi için tam benzeşme oranlarını göstermektedir. $\mathbf{RE}$ fonksiyonu giriş olarak alan isimlerini almakta düzenli ifadeler oluşturmaktadır. $\mathbf{ED}$ en az düzeltme mesafesi değeri ise Denklem (2)'de belirtildiği gibi atanmıştır. $\mathbf{KONTROL}$ fonksiyonu ise giriş olarak düzenli ifade, doküman ve en az düzeltme mesafesini almakta ve eğer eşleşme var ise 1, yok ise 0 değerini döndürmektedir. Eşleşme olması durumunda benzerlik oranı (b_i) ilgili alan ismine atanan etki değeri (e_i) olmaktır, aksi durumda ise ilk atanan 0 değeri olmaktadır.

Algoritma 4'de doküman tipi kararı için tasarlanan işlemler verilmiştir. Burada $\mathbf{s}$ eşik değerini, $\mathbf{k}$ ise dokümana atanan sınıfı belirtmektedir. Öncelikle dokümanın en çok benzediği doküman türü belirlenmektedir ($max(\mathbf{B})$). Eğer bu benzeşme değeri bu doküman türüne ait tam benzeşme değerine oranla belirli bir sınırı aşıyor ise sınıf ataması yapılmaktadır, aksi durumda harici bir doküman olduğu kararı verilmektedir. Bu çalışmada $\mathbf{s}$ eşik değeri deneysel olarak 0.25 olarak belirlenmiştir. Örneğin TC kimlik kartı için tam benzerlik oranı değeri 50 olduğu durumda, tipi belirlenmek istenilen bir dokümanın benzerlik oranlarının en yüksek değeri 12.5 veya üzerinde ise TC kimlik kartı ataması yapılabilmektedir. Bu şart doküman tipi belirlemede yüksek etki değerli alan isimlerinde eşleşme olmasını zorunlu kılarak sistemin karar vermede doğrulunu sağlayan önemli bir faktördür.

Algoritma 3: Benzeşme Oranı Tespiti

Girdi: $\mathbf{Dok}, \mathbf{ED}, \mathbf{E} = [e_1, e_2, \ldots, e_m], \mathbf{D} = [d_1, d_2, \ldots, d_m]$
Çıktı: $\mathbf{B} = [b_1, b_2, \ldots, b_n]_{ilk\_atama=0}, \mathbf{T} = [t_1, t_2, \ldots, t_n]_{ilk\_atama=0}$
for *each* d_i **do**
  for *each* d_{ij} **do**
    If $\mathbf{KONTROL}(\mathbf{RE}(d_{ij}), \mathbf{Dok}, \mathbf{ED}))$
      $b_i \leftarrow b_i + e_{ij}$
    end
    $t_i \leftarrow t_i + e_{ij}$
  end
end

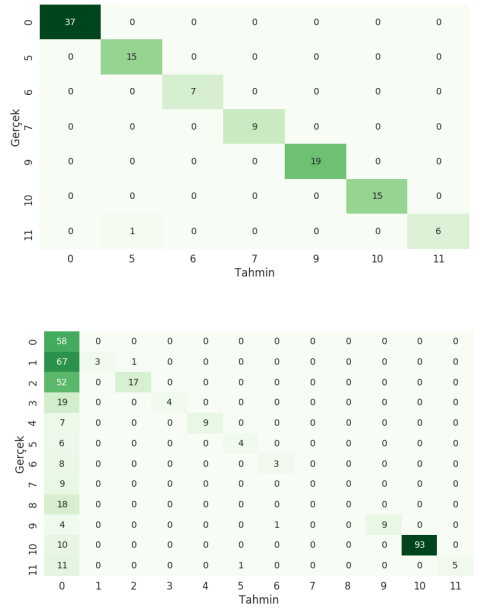
Algoritma 4: Karar Verme

Girdi: $\mathbf{B}, \mathbf{T}, \mathbf{s}$
Çıktı: $\mathbf{k}$
If *max*$(\mathbf{B}) \geq \mathbf{s} \times \mathbf{T}(arg\,max(\mathbf{B}))$
  $\mathbf{k} \leftarrow arg\,max(\mathbf{B})$
else
  $\mathbf{k} \leftarrow$ diğer
end

IV. SONUÇLAR

Bu çalışmada, Türkiye Cumhuriyeti'nde kullanılan resmi dokümanların kural tabanlı sınıflandırılması amaçlanmıştır. Literatürde benzer bir yaklaşımın yapılabilmesi için herhangi

bir veri setinin mevcut olmadığı görülmüştür. Söz konusu dokümanların kişiye özel olmasından dolayı geliştirme ve test esnasında temin edilmesi hukuksal anlamda bir takım prosedürlerin yerine getirilmesini gerektirmiştir. Bu sebeple kısıtlı olsa da gerçek veriler ile önerilen metotların testi sağlanmıştır. Hazırlanan veri setinde doğrudan dijital ortamda oluşturulan 108 adet doküman ve kamera veya tarayıcı aracılığı ile dijital ortama aktarılan 417 adet doküman bulunmaktadır. Bu 2 katagorideki veriler ayrı ayrı test edilerek doğrulama matrisleri Şekil 2'de sunulmuştur.



Şekil 2: Doğrulama matrisi (a) doğrudan dijital ortamda oluşturulan dokümanlar, (b) kamera veya tarayıcı aracılığıyla dijital ortama aktarılan dokümalar için. Burada 0 etiketi harici doküman tipini, diğerleri ise Bölüm I'de D ile belirtilen doküman tipi numaralarını temsil etmektedir.

D-5, D-6, D-7, D-9, D-10 ve D-11 tiplerine ait dokümanlara doğrudan dijital ortamda oluşturturulmuş formda karşılaşılabildiği için Şekil 2-a'da 6 adet doküman tipi sınırlandırması yapılmıştır. Test sonrasında doküman tiplerinin tamamının doğru tespit edildiği görülmektedir (genel başarım %100). Şunu özellikle belirtmek isteriz ki, bu sonuç veri çıkarımının sorunsuz olarak gerçekleştirilebildiği (OKT bağımlılığı olmadan) doküman tiplerinde önerilen metodun doğru karar verebildiğini ve doğru modelleme yeteneğine sahip olduğunu göstermektedir.

Doküman tipinin hatalı atanması yerine mevcut doküman tiplerinin haricinde "diğer" doküman tipi şeklinde atama yapabilabilmesi beklenmektedir. Örneğin, verilen bir TC kimlik kartı için eğer adli sicil kaydı ataması yapılırsa bütüncül sistemin doğru çalışması mümkün olamayacaktır. Bu kısıt ve gereklilikler ışığında Şekil 2-b incelendiğinde öncelikle test kümesinde bulunan 58 adet harici tipdeki dokümanın hatalı olarak herhangi bir sınıfa atanmadığı görülmektedir. Ayrıca tipi hatalı olarak tespit edilen dokümanların neredeyse hepsine yanlızca harici doküman tipi ataması yapılmıştır. Vurgulamak isteriz ki, bu durum karar verme algoritmasında çalışmanın geri kalan bölümlerinde harici dokümanlar için yanıltıcı bir etkinin mevcut olmadığını göstermektedir. Şekil 2-b incelendiğinde genel başarım %49.16'dır. Ancak, "harici" doküman sınıfına hatalı atamalar göz ardı edilirse %98.55 olduğu gözlemlenmektedir. Kamera veya tarayıcı aracılığıyla dijital ortama aktarılan dokümanların çözünürlük ve boyut gibi etkenlerden dolayı doğru okulabilmesi ağırlıklı olarak OKT kalitesine bağımlıdır. Bu sebeple, Şekil 2-b'de doküman tipinin belirlenmesinde hatalar olduğu gözlemlenmiştir.

Çalışmanın devamı olarak başarım arttırımı için ekibimiz tarafından evrişimli sinir ağları kullanılarak makina öğrenmesi tabanlı bir sistem halihazırda geliştirilmektedir ve nihai durumda mevcut çalışmada önerilen metot ile hibrit olarak çalışması planlanmaktadır. Bu çalışma, geliştirilmekte olan DATAMIN isimli KVKK uyum sürecine yardımcı olması hedeflenen endüstriyel bir ürünün ilk aşaması olan kural tabanlı resmi dokuman sınıflandırma yaklaşımını içermektedir. Yapılan kural tabanlı doküman sınıflandırma sonrasında belirlenen doküman tipine uygun makina öğrenmesi ve sözlük tabanlı hibrit metotlar kullanılarak veri çıkarımı ile ilişkilendirmesi planlanmaktadır.

BİLGİLENDİRME

Bu çalışma 3190083 numaralı 1051-TÜBİTAK Sanayi Ar-Ge Projeleri Destekleme Programı tarafından desteklenmiştir.

KAYNAKLAR

[1] M. Otto, "Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation – GDPR)," in *International and European Labour Law*, vol. 2014, pp. 958–981, Nomos Verlagsgesellschaft mbH & Co. KG, 2018.

[2] Türkiye Büyük Millet Meclisi, "6698 sayılı kişisel verileri koruma kanunu," 2016. https://www.mevzuat.gov.tr/MevzuatMetin/1.5.6698.pdf.

[3] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based Framework for Text Categorization," *Procedia Engineering*, vol. 69, pp. 1356–1364, 2014.

[4] K. Spärck Jones, "IDF term weighting and IR research lessons," *Journal of Documentation*, vol. 60, pp. 521–523, oct 2004.

[5] G. SahIn, "Turkish document classification based on Word2Vec and SVM classifier," in *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, may 2017.

[6] E.-H. S. Han and G. Karypis, "Centroid-based document classification: Analysis and experimental results," in *European conference on principles of data mining and knowledge discovery*, pp. 424–431, Springer, 2000.

[7] J.-Y. Yoo and D. Yang, "Classification scheme of unstructured text document using tf-idf and naive bayes classifier," *Advanced Science and Technology Letters*, vol. 111, no. 50, pp. 263–266, 2015.

[8] A. W. Harley, A. Ufkes, and K. G. Derpanis, "Evaluation of deep convolutional nets for document image classification and retrieval," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 991–995, IEEE, 2015.

[9] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for document image classification," in *2014 22nd International Conference on Pattern Recognition*, pp. 3168–3172, IEEE, 2014.

[10] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[11] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000.

[12] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, pp. 707–710, 1966.