Journal of the Faculty of Engineering and Architecture of Gazi University 31:4 (2016) 1085-1094

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Doküman dili tanıma için yeni bir öznitelik çıkarım yaklaşımı: İkili desenler

Yılmaz Kaya[1*], Ömer Faruk Ertuğrul[2]
[1]Siirt Üniversitesi, Bilgisayar Mühendisliği Bölümü, Kezer Kampüsü, Siirt, Türkiye
[2]Batman Üniversitesi, Elektrik-Elektronik Mühendisliği Bölümü, Batıraman Kampüsü, Batman, Türkiye

Ö N E Ç I K A N L A R

- Bu çalışmada yeni bir dil tanıma yaklaşımı önerilmiştir
- Metinlerin öznitelikleri bir boyutlu yerel ikili örüntüler (1B-YİÖ) ile çıkarılmıştır
- Çıkarılan öznitelikler yapay öğrenme metotları ile sınıflandırılmıştır

Makale Bilgileri
Geliş: 30.10.2015
Kabul: 05.04.2016

DOI:
10.17341/gummfd.45983

Anahtar Kelimeler:
Metin tabanlı dil tanıma, yerel ikili desenler, doğal dil işleme, öznitelik çıkarma

ÖZET

Doğal dil işlemenin önemli alt konularından biri olan dil tanıma (DT) bir dokümanın içeriğine göre yazıldığı dili belirleme işlemidir. Bu çalışmada, karakterlerin UTF-8 değerlerini birbirleri ile karşılaştırması sonucu elde edilen ikili desenler kullanarak yeni bir dil tanıma yaklaşımı önerilmiştir. Önerilen bu yöntemin başarısını test etmek amacıyla farklı sayıda dillerden oluşan metinler içeren dört veri kümesi kullanılmıştır. Önerilen yöntemde bir (1) boyutlu yerel ikili örüntüler (1B-YİD) ile dokümanlardan elde edilen öznitelikler farklı makine öğrenme yöntemleri ile sınıflandırılarak metinlerde DT işlemi gerçekleştirilmiştir. Dört farklı veri kümesi için elde edilen DT başarı oranları sırası ile %86,20, %92,75, %100 ve %89,77 olarak gözlenmiştir. Elde edilen sonuçlara göre önerilen öznitelik çıkarım yönteminin dil tanıma için önemli örüntüler sağladığı görülmüştür.

A novel feature extraction approach for text-based language identification: Binary patterns

H I G H L I G H T S

- In this paper, a novel language identification approach was proposed
- The features of the text messages were extracted by one-dimensional local binary patterns (1D-LBP)
- Extracted features were classified by machine learning methods

Article Info
Received: 30.10.2015
Accepted: 05.04.2016

DOI:
10.17341/gummfd.45983

Keywords:
Text-based language identification, one dimensional local binary patterns, natural language processing, feature extraction

ABSTRACT

Language identification (LI), which is a major task in natural language processing, is the process of determining the language from a given content. In this paper, a novel approach, which is based on the probability of the use of the characters that have the similar orders with respect to their UTF-8 values, was proposed. In order to evaluate and validate the proposed approach, four datasets, which contain texts in different numbers of languages, were employed. In the proposed approach, the features that were exacted by one-dimensional local binary pattern (1D-LBP) method were classified by various machine learning methods. Achieved LI accuracies in each of four employed datasets were 86.20%, 92.75%, 100% and 89.77%, respectively. The results showed that the proposed approach yields high success rates and it is an efficient way of language identification.

* Sorumlu Yazar/Corresponding author: yilmazkaya1977@gmail.com / Tel: +90 484 212 1111

1. GİRİŞ (INTRODUCTION)

Son yıllarda artan web sayfaları nedeniyle bu sayfaların içeriklerinin tanımlanması veya bu sayfalardan bilgi çıkarımı için yeni tekniklere ihtiyaç duyulmaktadır [1]. Bilgi çıkarımı sürecinde içeriğin dilinin tanımlanması önemli bir aşamadır. Dil tanıma (DT) bir dokümanın içeriğini kullanarak dokümanın hangi dile (örneğin İngilizce, Türkçe, Arapça, vb.) ait olduğunun otomatik olarak tespitidir [2]. DT için literatürde dilbilimsel (linguistik) veya istatistiksel tabanlı farklı yaklaşımlar kullanılmıştır. Dilbilimsel yaklaşımlar bir dile ait özel bir kelime veya karakteri arayan ve endeksleyen metotlar olup dillere ait kuralları temel alarak çalışmaktadırlar. İstatistiksel yaklaşımlar da ise dili oluşturan kelime veya karakterlerin frekans ve dağılımlarına bağlıdır. İstatistiksel yaklaşımlar içerik-bağımsız yöntemler dilbilimsel yöntemlere kıyasen yeterli bilgi vermemektedir. Daha ziyade dilleri matematiksel olarak modellemek başarıyla için kullanılan bu yaklaşımların en büyük dezavantajı benzer dillerde ayırt etme başarısının düşük olmasıdır [2]. Öte yandan DT dokümana ait kelime veya karakter boyutunda elde edilen özellikleri temel alan bir metin sınıflandırma problemi olarak ta ifade edilebilir [3]. Literatürde web tabanlı dokümanları kullanılarak bilgi çıkarımı [4], konuşulan dili modelleme [5], çoklu dil çeviri sistemleri [6], spam tespiti [7], doküman sınıflama [1, 8, 9], dijital kütüphane oluşturma [10], metin özetleme [11], otomatik soru-cevap sistemleri ve çeviri sistemleri [12] DT ile ilgili gerçekleştirilen uygulamaların bir kısmıdır. DT amacıyla tekil karakter kombinasyonları [13], kısa kelime [14], n-gram [15] ve ASCII/Unicode karakter frekans vektörleri gibi çeşitli öznitelik çıkarım yöntemleri kullanılmıştır. Genelde karakter seviyesinden çıkarılan özellikler kullanılarak elde edilen sonuçlar kelime düzeyindeki özellikler kullanılarak elde edilen sonuçlara kıyasen daha kararlıdırlar [16, 17, 18] ve karakter seviyesinde öznitelik çıkaran n-gram yönteminin diğer öznitelik çıkarım yöntemlerine kıyasen daha başarılı olduğu raporlanmıştır [19]. Ancak bu yöntemde çok fazla öznitelik çıkarılması nedeniyle öznitelik uzayı büyümekte ve bu da yüksek işlem yük ve hafıza ihtiyacı gibi sorunlara neden olmaktadır. Bu sebeple genelde bu yöntemle beraber öznitelik seçim yöntemleri kullanılmaktadır. Bu çalışmada ise karakterlerin UTF-8 değerlerinin ikili karşılaştırmalar sonucu elde edilen desenleri kullanan yeni bir öznitelik çıkarım yöntemi olarak bir boyutlu yerel ikili desenler (1B-YİD) metodu önerilmiştir. Bu çalışmada görüntüdeki yerel değişimleri kullanarak görüntülerden öznitelik çıkarmak için kullanılan yerel ikili desenler (YİD) metodu tek boyutlu hale getirilerek metin madenciliğinde bir öznitelik çıkarım metodu olarak kullanılmıştır [20]. Bu yöntemde öncellikle tüm karakterler Unicode değerlerine çevrilmekte ve daha sonra her bir karaktere karşılık gelen değer kendi komşularıyla karşılaştırılmaktadır. Karşılaştırmada eğer söz konusu karakterin Unicode karşılığı komşu karakterlerin Unicode karşılığından büyük ise o komşu karakter için 1 aksi halde 0 değeri atanarak bir ikili dizge elde edilmektedir. Bu ikili dizgelerin onlu karşılığı karşılaştırılan karakterin yeni değeri olarak alınmaktadır. Bu şekilde tüm karakterler için elde edilen yeni değerler 1B-YİD sinyalini oluşturmakta ve bu sinyale ait histogram vektörü öznitelik vektörü olarak kullanılmaktadır. 1B-YİD yönteminde faklı mikro-makro desenlerin tespit edilmesi için, komşuları belirleyen P, α ve β gibi üç parametre ile tespit edilebilmektedir. Bu çalışmada birçok alanda başarıyla kullanılan YİD metodunun kazanımlarının DT alanında kullanımı amaçlanmıştır. Önerilen yöntemi test etmek için farklı şekillerde oluşturulmuş dört veri kümesi kullanılmıştır. Elde edilen özellikler ise yapay sinir ağları (YSA), destek vektör makineleri (DVM) ve çok terimli Naive Bayes (ÇTNB) gibi farklı sınıflandırma metotları kullanılmıştır. Elde edilen sonuçlara önerilen yöntemin dil tanımada başarıyla kullanılabileceği görülmüştür.

2. YAPILAN ÇALIŞMALAR (RELATED WORKS)

Doğal dil işleme uygulamaları için DT önemli bir yer tutmaktadır. Bu sebeple DT süreçlerinin öznitelik çıkarımı, bilgi çıkarımı ve sınıflandırma basamaklarına yönelik önemli bir ilgi gösterilmiştir [21]. Bu amaçla en yaygın kullanılan öznitelik çıkarım yöntemleri olarak n-gram ve kısa kelimeler göze çarpmaktadır [22, 23]. N-gram ya da kısa kelimeler gibi istatiksel modellere ilaveten ayrıştırılmak istenen dillere ait özel karakterler veya dilin karakteristik yapıları da bu amaçla kullanılmaktadır [13, 24]. Sınıflandırma amacıyla Markov modelleri [3, 25], entropi tabanlı metotları [26], Gaussian karışımlı modelleri [27], karar ağaçları [28], YSA [29], DVM [30], melez modelleri [3, 31], kNN ve regresyon modelleri [32, 33] uygulanan makine öğrenmesi yöntemlerinden sadece bir kısmıdır. DT amacıyla yapılan çalışmaların bir kısmında kullanılan metotlar ve elde edilen başarı oranlarını Tablo 1'de özetlenmiştir. Yapılan çalışmalara bakıldığında, DT için yapılan çalışmaların önemli bir kısmının öznitelik seçim tabanlı olduğu görülmektedir. Ancak, Yavanoğlu ve Sağıroğlu'nun da belirttiği gibi literatürdeki mevcut yöntemlerin yetersiz olup bu anlamda DT için yeni metotlara ihtiyaç duyulmaktadır [35].

3. KULLANILAN VERİ KÜMELERİ (UTILIZED DATASETS)

Bu çalışmada önerilen yaklaşımın DT başarısını test etmek için aşağıda özellikleri verilen 4 farklı veri kümesi kullanıldı. *1. veri kümesi* (VK-1): Bu veri kümesi aşk, iktidar, barış, bilgisayar, bilişim, teknoloji, insanlık, aile, mutluluk, kanser, spor, uzay, para vb. 110 farklı kelimenin Türkçe Wikipedia kullanılarak elde edilen anlamları temel alınarak oluşturuldu. Daha sonra elde edilen bu metinler Google çeviri sistemi ile Fransızca, Almanca, Hollandaca, İtalyanca, Portekizce, İspanyolca, İsveççe, Çince ve İngilizce dillerine çevrildi. Bu veri kümesinde 10 farklı dilde 500-1000 karakter boyutlarında toplamda 1100 metin elde edilmiş oldu.

Tablo 1. Literatürde kullanılan yöntemler ve elde edilen başarı oranları
(Employed methods and obtained results in the literature)

Referans	Veri Kümesindeki Dil Sayısı	Öznitelik Çıkarım Yöntemi	Sınıflandırma Yöntemi	Başarı Oranı (%)
[21]	9	Harf frekansı	Ağırlık merkezi ve ters sınıf frekansı	98
[23]	18	1-gram	Bağıl entropi	78,2-99,4
		2-gram		90,2-100
[31]	13	2-gram	Linguini (vektorel	68,8-100
		3-gram	mesafe tabanlı bir	79,5-100
		4-gram	yöntem)	83,6-100
		5-gram		81,4-99,9
		Kısa kelimeler		61,3-100
		Kısa kelimeler + 3-gram		83,8-100
		Kısa kelimeler + 4-gram		84,9-100
[33]	4	Harf frekansı	YSA ve bulanık mantık	99,75
[34]	15	n-gram	Ağırlık Merkezi Tabanlı	93-93,5
			DVM	61-85
			YSA	77-91
			k-ortalamalar	64-76,4
			Bulanık C ortalamaları	77-86,6
[35]	15	Birleşim tespit yöntemi	YSA	90-98

2. veri kümesi (VK-2): Bu veri kümesi BBC web sitesinden (www.bbc.com) İngilizce, Almanca ve Fransızca dillerinin her biri için spor, sanat, teknoloji, güncel haberler kategorilerinden rasgele 100'er haber ve Türkçe haber sitelerinden de benzer kategoriler için 100 metin olmak üzere toplamda 400 metinden oluşmaktadır. *3. veri kümesi* (VK-3): Bu veri kümesinde 4 farklı dilde yayınlanmış olan İnsan Hakları Evrensel Belgesi'nden (http://www.cis.hut.fi/research/cog/data/udhr/) elde edilen toplamda 24 farklı metinden oluşmaktadır. *4. veri kümesi* (VK-4): Bu veri kümesi Baldwin ve Lui tarafından wikipedia adıyla oluşturulmuş birçok çalışmada kıyaslama için kullanılan çok dilli bir veri kümesidir [36]. Bu çalışmada bu veri setinde bulunan 500 karakterin altındaki metinler atıldıktan sonra geriye kalan 25 dile ait metinler kullanılmıştır. VK-4'ün doküman dağılımları Tablo 2'de verilmiştir. Tüm veri kümelerinde bulunan metinlerin içeriğinde bulunan özel isimler, noktalama işaretleri, boşluklar ve özel işaretler metinlerden çıkarılmış ve metinler UTF-8 formatına kullanılmıştır.

4. METOT (METHOD)

4.1. Bir Boyutlu Yerel İkili Desenler Yöntemi
(One Dimensional Local Binary Patterns)

1B-YİD yöntemi metinlerden yeni öznitelik çıkarımı için görüntü işlemede yaygın bir şekilde kullanılan YİD metodundan geliştirilmiştir [37]. 1B-YİD yöntemi işleyiş olarak görüntü işlemede kullanılan YİD yöntemi ile benzerlik göstermektedir. Ancak, 1B-YİD yöntemi görüntü yerine zaman serisi şeklinde dizilmiş tek boyutlu sinyallere ve metinlere uygulanmıştır [37, 38]. 1B-YİD yönteminde sinyaldeki her bir değer için ile komşuları arasında yapılan karşılaştırmalar sonucu ikili dizgeler üretilir. Elde edilen bu dizgelerin onluk karşılıkları sinyalin 1B-YİD dönüşümü olarak ifade edilebilir. İkili karşılaştırmalar Eş. 1'de verilen şekilde hesaplanmaktadır.

$$t = P_i - Pc$$

$$1B-YIO(x) = \sum_{i=0}^{P} İşaret(t)2^{i-1}$$

$$İşaret = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (1)$$

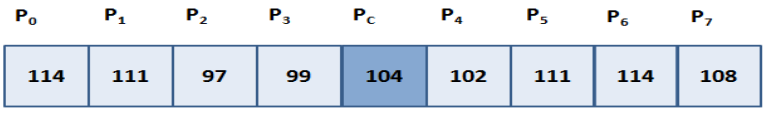
Burada P_i ve P_c sırasıyla ele alınan komşular ve karşılaştırılan merkez değeri belirtir. 1B-YİD yöntemi P, α ve β parametreleri kullanılarak optimize edilmektedir. P merkez noktanın sağından ve solundan alınacak toplam komşu sayısını belirtmektedir. α merkez nokta ile alınacak ilk komşular arasındaki mesafeyi, β ise alınan komşular arasındaki mesafeyi göstermektedir.

$$\alpha = |P_c - P_i| \text{ ve } \beta = |P_i - P_j| \quad (2)$$

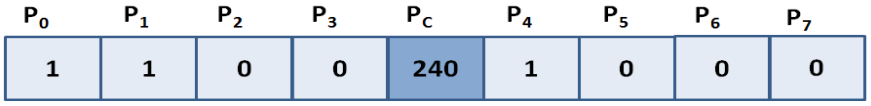
1B-YİD yönteminde sinyaldeki her bir değer kendi komşuları ile karşılaştırılmaktadır. Sinyaldeki her bir değerin öncesinden ve sonrasından (sağından ve solundan) *P/2* kadar değer o değerin komşusu alınır. Örneğin *P=8* olması durumunda Şekil 1'de gösterildiği gibi her değerin (P_c) öncesinde 4 komşu (P_0, P_1, P_2, P_3) ve sonrasında 4 komşu (P_4, P_5, P_6, P_7) alınır. Daha sonra merkez değer (P_c) ile bu değerin komşuları (P_i) karşılaştırılıp Eş. 1'e göre her bir karşılaştırma için bir ikili değer elde edilir. Karşılaştırmalarda eğer P_i değeri P_c'den büyük ve eşit ise 1, diğer durumlarda 0 alınır ve bu karşılaştırmalar sonucunda 1B-YİD kodu oluşur. Şekil 1'de gösterilen sinyal parçacığı için elde edilen karşılaştırma sonuçları (ikili değerler) ve bu değerlerin onluk karşılığı Şekil 2'de gösterilmiştir.

Tablo 2. VK-4 için metin dağılımları (Distribution of the languages in VK4)

No	Dil	Metin Sayısı	No	Dil	Metin Sayısı
1	Çince	353	14	İzlandaca	42
2	Japonca	196	15	Zuluca	40
3	Afrikanca	187	16	İngilizce	40
4	Tagalogça	126	17	Hollandaca	40
5	Danimarkaca	91	18	İspanyolca	38
6	İsveççe	79	19	Hintçe	35
7	Fince	75	20	Almanca	34
8	Malezyaca	69	21	Makedonca	34
9	İtalyanca	57	22	Galce	32
10	Peştuca	53	23	Hırvatça	31
11	Estonyaca	48	24	Catalanca	30
12	Fransızca	46	25	Tayça	28
13	Farsça	45			



Şekil 1. Sinyal üzerindeki örnek bir değer (A sample point on signal)



Şekil 2. P_c'nin P_i ile karşılaştırılması (Comparison of P_c with P_i)

Her bir P_c için elde edilen ikili kodların onlu karşılığı o P_c noktasının etrafındaki yerel bilgileri ifade etmektedir. Yukarıda ifade edilen aşamalar sinyal üzerindeki her bir değer için gerçekleştirilmekte ve sinyalin 1B-YİD karşılığı elde edilmektedir. Sinyalin 1B-YİD karşılığındaki her değerin frekansı farklı bir deseni ifade etmektedir. *P=8* olması durumunda 1B-YİD sinyali üzerindeki tüm değerler 0 ile 255 arasındaki değişim göstermektedir. Elde edilen 1B-YİD sinyalinin tümünü içeren YİD^{Tumu}, tekdüze olanları (YİD^{u2}) ve tekdüze olmayan (YİD^{nu2}) şeklinde kullanılabilmektedir. Tekdüze özellikler ikili dizgideki bit geçiş (0'dan 1'e veya 1'den 0'a) sayısı en fazla 2 olan özniteliklerdir. Örneğin 11100001 deseninde geçiş sayısı 2 olduğu için tekdüze özelliktir. 11110101 deseninde ise geçiş sayısı 4 olduğundan tek düze özellik değildir. *P=8* olması durumunda sinyalden çıkarılacak özellik sayısı YİD^{Tumu} için 256 iken YİD^{u2} için 58'dir.

4.2. Önerilen Metot (Proposed Method)

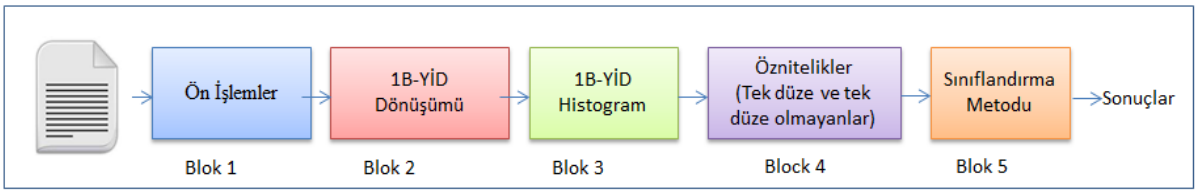
Bu çalışmada DT için önerilmiş olan çalışmalardan tümüyle farklı yeni bir yaklaşım önerilmiştir. Önerilen yöntem karakter seviyesinde istatistiksel bir yaklaşım olup blok diyagramı Şekil 3'te verilmiştir.

Blok 1: Bu blokta metin içende geçen boşluklar, noktalama işaretleri, yeni satır gibi özel karakterler atılır ve metin Unicode'lara dönüştürülür. Unicode'lerden oluşan yeni dizi bir boyutlu sinyal olarak ele alınır. Örneğin metin olarak: "*Doküman dili tanıma için yeni bir öznitelik çıkarım yaklaşımı: İkili Desenler*" alındığında öncelikle metin içendeki istenilmeyen karakterler atılır ve geriye kalan metin: "*Doküman dili tanıma için yeni bir öznitelik çıkarım yaklaşımı İkili Desenler*" olup, bu metin UTF-8 kodlarına dönüştürülmesi sonucunda aşağıdaki bir sinyal aşağıdaki gibi olur.

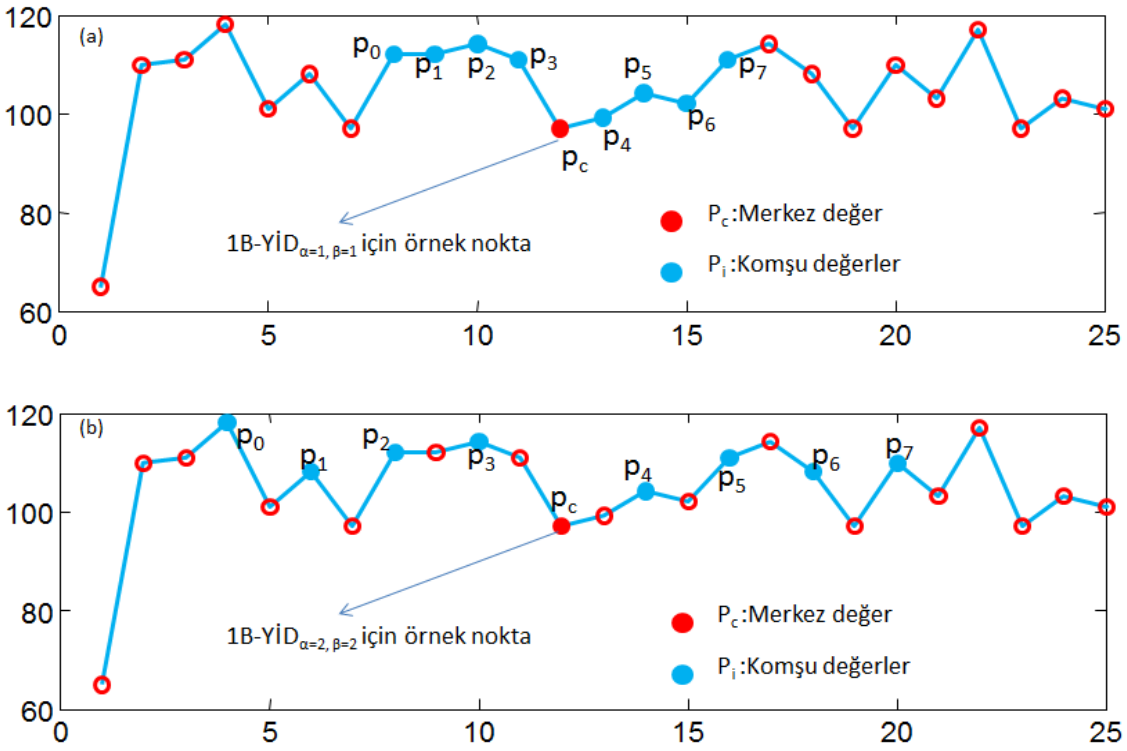
"68, 111, 107, 252, 109, 97, 110, 100, 105, 108, 105, 116, 97, 110, 253, 109, 97, 105, 231, 105, 110, 121, 101, 110, 105, 98, 105, 114, 246, 122, 110, 105, 116, 101, 108, 105, 107, 231, 253, 107, 97, 114, 253, 109, 121, 97, 107, 108, 97, 254, 253, 109, 253, 221, 107, 105, 108, 105, 68, 101, 115, 101, 110, 108, 101, 114"

Blok 2: Elde edilen UTF-8 kodlarını içeren sinyali 1B-YİD metodu ile yeni bir düzleme taşınmaktadır. Bu düzlemde her bir değerin frekansı farklı bir örüntü tanımlamaktadır. Farklı *P*, α ve β parametreleri alındığında farklı örnekler Şekil 4'te gösterildiği gibi elde edilmektedir. Şekil 4'te görüldüğü gibi 1B-YİD parametrelerinin farklı değerlerine göre aynı sinyal parçası için farklı örüntüler elde edilmektedir.

Blok 3: Bu blokta metne karşılık gelen 1B-YİD sinyaline ait histogram elde edilmektedir. Histogramın her bir değeri bir örüntü veya öznitelik olarak değerlendirilir. *P=8* olması durumunda 256 örüntü bulunmaktadır.



Şekil 3. Önerilen yönteme ait blok diyagram (The block diagram of the proposed method)



Şekil 4. Metine ait örnek bir sinyal bölümü (A sample signal from the message)

Blok 4: Elde edilen özellikler tekdüze ve tek düze olmayan özellikler şeklinde ayrıştırılmış ve farklı veri kümeleri olarak kullanılmıştır.

Blok 5: Bu aşamada elde edilen öznitelikler YSA, DVM ve ÇTNB gibi farklı sınıflandırma metotları ile 10 katlı çapraz doğrulama yöntemine göre sınıflandırılmıştır. Yapılan çalışmada kullanılan başarı ölçütleri aşağıda gösterilmiştir (Eş. 3-Eş. 8) [38].

$$Başarı\ Oranı\ (\%) = \frac{N_D}{N_Y+N_D} * 100\% \quad (3)$$

$$Doğru\ Tahmin\ Oranı\ (DTO) = \frac{N_D}{N_Y+N_D} \quad (4)$$

$$Yanlış\ Tahmin\ Oranı\ (YTO) = \frac{N_Y}{N_Y+N_D} \quad (5)$$

$$Geri\ Çağırım = \frac{N_{DP}}{N_{YN}+N_{DP}} \quad (6)$$

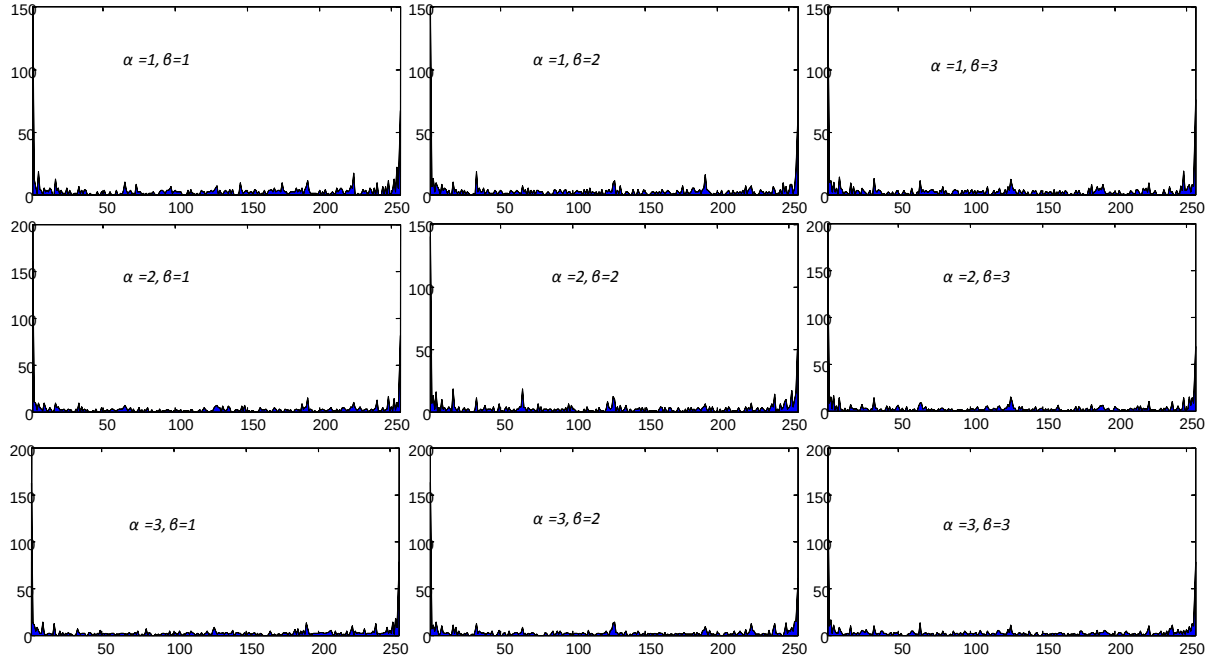
$$Duyarlılık = \frac{N_{DP}}{N_{YP}+N_{DP}} \quad (7)$$

$$F-Ölçütü = \frac{2xN_{DP}}{2xN_{DP}+N_{YP}+N_{YN}} \quad (8)$$

Başarı, doğru ve yanlış tahmin oranları genel her bir dil için tek tek hesaplanabileceği gibi tüm diller için genel bir DT başarısı da hesaplanabilmektedir. Öte yandan geri çağırım, duyarlılık ve f-ölçütü göstergeleri ise her bir dil için tek tek hesaplanabilmektedir. Burada N_D doğru sınıflandırılan metin sayısı; N_Y ise yanlış hesaplanan metin sayısını göstermektedir. N_{DP} söz konusu dile ait doğru sınıflandırılan metin sayısını, N_{FP} söz konusu dile ait olduğu halde yanlış sınıflandırılan dil sayısı ve N_{FN} ise söz konusu dile ait olmayan yanlış sınıflandırılmış metin sayısını göstermektedir.

5. SONUÇLAR VE TARTIŞMALAR (RESULTS AND DISCUSSIONS)

Bu çalışmada DT amacıyla karakterlerin UTF-8 değerlerini kullanarak yeni bir yaklaşım önerilmiştir. Önerilen 1B – YİD $_{P,\alpha,\beta}$ metotta P, α ve β parametreleri kullanılarak farklı mikro-makro örüntüler tespit edilebilmektedir. Bir örnek



Şekil 5. α ve β parametrelerine göre örüntülerin dağılımı (Distribution of patterns according to α and β parameters)

metin için bu parametrelerin farklı değerlerine göre elde edilen örüntülerin histogram dağılımı için Şekil 5'te verilmiştir. Şekil 5'te görüldüğü gibi α ve β parametrelerin farklı değerlerine göre elde edilen histogramın değiştiği görülmektedir. Bu parametrelerin farklı değerleri ile veri kümelerinden elde edilen örüntüler ÇTNB ile sınıflandırılmış ve elde edilen başarı oranları Tablo 3-5'te verilmiştir.

Tablo 3. VK-1 için başarı oranları (%)
(Obtained accuracies in VK1 (%))

	$\alpha=1$	$\alpha=2$	$\alpha=3$
$\beta=1$	86,20	69,50	61,20
$\beta=2$	69,50	58,20	55,30
$\beta=3$	57,30	53,90	45,90

Tablo 4. VK-2 için başarı oranları (%)
(Obtained accuracies in VK2 (%))

	$\alpha=1$	$\alpha=2$	$\alpha=3$
$\beta=1$	88,25	72,00	61,25
$\beta=2$	89,25	79,75	73,50
$\beta=3$	92,75	72,00	77,00

Tablo 5. VK-4 için başarı oranları (%)
(Obtained accuracies in VK4 (%))

	$\alpha=1$	$\alpha=2$	$\alpha=3$
$\beta=1$	89,77	82,20	80,47
$\beta=2$	81,55	67,82	66,36
$\beta=3$	75,22	63,49	58,51

Tablo 3-5'te görüldüğü gibi VK-1 ve VK-4 için α ve β parametre değerlerinin artması ile başarı oranların düştüğü görülmektedir. Ancak bu durum VK-2 için gözlenmemiştir. Maalesef önerilen yaklaşımda en uygun α ve β parametreleri ile sınıflandırma başarısı arasında anlamlı ve sürekli bir ilişki gözlenememiştir. VK-1'de 10, VK-2'de 4 ve VK-4'te ise 25 dil mevcut olduğu göz önüne alındığında, veri kümelerinde bulunan dil sayısı ile ilişki olabileceği tahmin edilmektedir. Genel olarak ise α ve β parametrelerin uygun değerleri deneme-yanılma sonucunda karar verilmelidir. Elde edilen sınıflandırma doğrulukları incelendiğinde $Y\dot{I}D^{Tumu}_{P,\alpha=1,\beta=1}$ ile VK-1 için %86,20, VK-4 için %89,77 başarı oranları VK-2 için ise $Y\dot{I}D^{Tumu}_{P,\alpha=1,\beta=3}$ ile %92,25 başarı oranı elde edilmiştir. VK-3 için %100 oranında başarı elde edilmiştir. Bu veri kümesinde yüksek başarı oranı elde edilmesinin nedeni olarak bu veri kümesindeki metinlerin karakter sayısının çok yüksek olduğunu düşünmekteyiz. VK-1 ve VK-4 için elde edilen duyarlılık, geri çağırım ve f-ölçütü performans ölçütleri Tablo 6-7'de verilmiştir. Tablo 6'dan de görüleceği üzere VK-1 için en yüksek tanıma başarısı Almanca ve İngilizce için elde edilmiştir. Bu dillere doğru sınıflandırılmış pozitif örnek sayısı oranları 0,97 ve 0,95 olarak gözlenmiştir. Öte yandan Tablo 5 için en ilginç bilgi Çince için elde edilen geri çağırım ve F-ölçütünün yüksek oranlarda gözlenmesi ve bu veri kümesi için yanlış sınıflandırılan hiçbir dokümanın Çinceye benzetilmemiştir. Tablo 7'den görüldüğü gibi VK-4'te Zuluca, Makedonca, Hintçe ve Tayca diğer dillerden kesin olarak ayrışmaktadır. Diğer diller için kabul edilebilir başarı oranları elde edilmiştir. En düşük başarı oranları Peştuca, İsveççe, İzlandaca ve Katalanca için elde edilmiştir. Metin uzunluklarının önerilen yöntemin başarısına etkisini test etmek amacıyla VK-1'de 100, 200 ve 300 karakter uzunluklarındaki metinler kullanılarak olarak denemeler gerçekleştirilmiş ve elde edilen başarı oranları Tablo 8'de verilmiştir.

Tablo 6. VK-1 seti için Performans Ölçütleri ($Y\dot{I}D^{Tümü}_{P=8,\alpha=1,\beta=1}$, ÇTNB)
(Obtained performance measures in VK1 ($LBP^{all}_{P=8,\alpha=1,\beta=1}$, MNB))

Dil	DTO	YTO	Duyarlılık	Geri Çağırım	F-ölçütü
Almanca	0,970	0,006	0,912	0,930	0,921
İngilizce	0,950	0,010	0,875	0,840	0,857
İtalyanca	0,900	0,021	0,850	0,850	0,850
Fransızca	0,900	0,022	0,842	0,850	0,846
Türkçe	0,890	0,011	0,923	0,960	0,941
Hollandaca	0,890	0,014	0,908	0,890	0,899
Çince	0,800	0,000	0,990	1,000	0,995
İsveççe	0,780	0,018	0,802	0,810	0,806
İspanyolca	0,780	0,032	0,758	0,720	0,738
Portekizce	0,750	0,018	0,755	0,770	0,762
Ortalamalar	0,860	0,017	0,861	0,862	0,862

Tablo 7. VK-4 için Performans Ölçütleri ($Y\dot{I}D^{Tümü}_{P=8,\alpha=1,\beta=1}$, ÇTNB)
(Obtained performance measures in VK4 ($LBP^{all}_{P=8,\alpha=1,\beta=1}$, NB))

Dil	DTO	YTO	Duyarlılık	Geri Çağırım	F-ölçütü
Zuluca	1,000	0,000	1,000	1,000	1,000
Hintçe	1,000	0,000	1,000	1,000	1,000
Makedonca	1,000	0,000	1,000	1,000	1,000
Tayça	1,000	0,000	1,000	1,000	1,000
İtalyanca	1,000	0,003	0,905	1,000	0,950
Afrikaca	0,979	0,006	0,948	0,979	0,963
Farsça	0,978	0,005	0,830	0,978	0,898
Fince	0,973	0,003	0,936	0,973	0,954
Almanca	0,971	0,002	0,917	0,971	0,943
Japonca	0,969	0,007	0,945	0,969	0,957
Fransızca	0,957	0,011	0,688	0,957	0,800
Tagalogça	0,944	0,002	0,975	0,944	0,960
Malezyaca	0,942	0,001	0,970	0,942	0,956
Çince	0,909	0,022	0,907	0,909	0,908
Hollandaca	0,900	0,001	0,947	0,900	0,923
Galce	0,875	0,001	0,933	0,875	0,903
Estonca	0,875	0,010	0,700	0,875	0,778
İngilizce	0,875	0,009	0,686	0,875	0,769
İspanyolca	0,842	0,005	0,780	0,842	0,810
Hırvatça	0,839	0,002	0,897	0,839	0,867
Danimarkaca	0,725	0,008	0,825	0,725	0,772
Katalanca	0,700	0,003	0,778	0,700	0,737
İzlandaca	0,667	0,001	0,933	0,667	0,778
İsveççe	0,608	0,005	0,842	0,608	0,706
Peştuca	0,604	0,003	0,842	0,604	0,703
Ortalamalar	0,898	0,008	0,901	0,898	0,896

Tablo 8. Farklı uzunlukta metinler için başarı oranları ($Y\dot{I}D^{Tümü}_{P=8,\alpha=1,\beta=1}$, 10 dil)
(Obtained accuracies in different text lengths ($LBP^{all}_{P=8,\alpha=1,\beta=1}$, 10 languages))

Karakter Uzunluğu	100	200	300
DT Başarı (%)	55,10	63,20	77,4

Literatürde de belirtildiği ve tablodan görüldüğü gibi DT'da başarı metin uzunluklarıyla direkt olarak ilişkilidir [22, 33]. Diğer bir deyişle, dil sınıflandırma başarısı metin karakter uzunluklarına bağlıdır. Tablo 8'den görüldüğü gibi önerilen yaklaşım ile kabul edilebilir başarı elde edilmesi için metin uzunluklarının en az 300 karakter olması gerekmektedir. Ayrıca VK-1 için 1B-YİD ile elde edilen tüm öznitelikler yerine sadece tekdüze özniteliklerin kullanılması ile elde edilen DT başarı oranları Tablo 9'da verilmiştir. Tüm öznitelikler yerine tekdüze öznitelikleri kullanmak öznitelik sayısı 256'dan 58'e indirmekte ve işlemsel yük azalmakta, ancak Tablo 9'dan da görüleceği gibi başarı oranını düşürmektedir. Bu sebeple önerilen öznitelik çıkarım yöntemi ile elde edilen tüm özniteliklerin kullanılması

sadece tekdüze özniteliklerin kullanılmasından daha uygun olacaktır. Ayrıca önerilen yöntemin başarısının ÇTNB bağlı olmadığını göstermek amacıyla $Y\dot{I}D^{tümü}_{P=8,\alpha=1,\beta=1}$ ile elde edilen öznitelikler ayrıca fonksiyonel karar ağaçları (FKA), doğrusal sınıflandırıcı (DS), YSA ve DVM yöntemleriyle sınıflandırılmış ve elde edilen başarı oranları Tablo 10'da verilmiştir. Sınıflandırma işlemi için bir veri madenciliği yazılımı olan, önişleme, özellik seçimi, sınıflandırma, kümeleme ve karar kuralları metotlarını içeren WEKA paket program kullanılmıştır [39].

Tablo 9. VK-1 için tekdüze öznitelikler ile elde edilen başarı oranları ($Y\dot{I}D^{u2}_{P=8,\alpha,\beta}$, 10 dil)
(Obtained accuracies in uniform features of VK1 ($LBP^{u2}_{P=8,\alpha,\beta}$, 10 languages))

	$\alpha=1$	$\alpha=2$	$\alpha=3$
$\beta=1$	68,70	67,40	62,20
$\beta=2$	64,80	59,10	53,20
$\beta=3$	53,70	56,60	49,80

Tablo 10: Sınıflandırma Başarı oranları ($Y\dot{I}D^{Tümü}_{P=8,\alpha=1,\beta=1}$)
(Obtained classification accuracies ($LBP^{all}_{P=8,\alpha=1,\beta=1}$)

Sınıflandırma Metodu	VK-1	VK-4
YSA	86,20	72,74
DVM	86,20	78,42
FKA	76,44	82,80
DS	85,80	89,72
ÇTNB	86,20	89,77

Tablodan görüleceği üzere VK-1 için en yüksek başarı oranı %86,20 olarak YSA, DVM ve ÇTNB ile elde edilmiştir. VK-4 için ise en yüksek başarı oranı %89,77 olup ÇTNB ile elde edilmiştir. Sonuç olarak 1B-YİD ile elde edilen öznitelikler kullanılarak farklı makine öğrenmesi yöntemleri ile kabul edilebilir başarı oranları elde edilmiştir. Çaprazlama DT sistemlerin geliştirilmesinde önemli bir konudur ve farklı veri kümelerini birinin eğitim diğerlerinin ise test amacıyla kullanılmasına dayanmaktadır. Tablo 11'de İngilizce, Fransızca, Almanca ve Türkçe için VK-1 ve VK-2 kullanılarak elde edilmiş çapraz başarı oranları verilmiştir.

Tablo 11. VK-1 ve VK-2 için çaprazlama başarı oranları
(Obtained accuracies in cross-domain of VK-1 and VK2)

Eğitim	Test	Başarı %
VK-1	VK-2	93,75
VK-2	VK-1	81,50

Elde edilen sonuçlara göre önerilen yöntemin çaprazlama durumunda bile önemli başarılar sağladığı görülmüştür (Tablo 11). Önerilen yöntemi DT için karakter n-gram yöntemi ile karşılaştırılmıştır. VK-2'den 2-gram ile çıkarılan özellikler ÇTNB ile sınıflandırıldığında %91,25 başarı oranı elde edilmiştir. Önerilen yaklaşımda elde edilen başarı oranının %92,77 olduğu göz önüne alındığında n-gram yöntemiyle daha yüksek öznitelik uzayı oluşturulmasına rağmen önerilen yaklaşıma göre daha başarısız olmuştur. Ayrıca Baldwin ve Lui yaptıkları çalışmada VK-4'den 2-gram ile elde edilen özniteliklerin kosinüs benzerliği ve 1 en yakın komşuluk yöntemleriyle sınıflandırılması elde ettikleri en yüksek duyarlılık, geri çağırım ve f-ölçütü değerleri ve başarı oranı 0,740, 0,646, 0,671 ve %86,9 olarak verilmiştir [36]. Ayrıca aynı çalışmada (1-5)-gram ile çıkarılan özniteliklerden etkin özelliklerin seçilmesi sonucunda ulaşılan en yüksek başarı yüksek duyarlılık, geri çağırım ve f-ölçütü değerleri ve başarı oranı 0,789, 0,718, 0,729, %90,2 olarak raporlanmıştır [36]. Önerilen yaklaşımda elde edilen başarı ölçütleri daha yüksektir. Önerilen yaklaşımın en önemli eksiği en uygun *P*, α ve β parametre değerlerinin tespitine ilişkindir. Bu değerler n-gram yönteminde en uygun "n" değerinin atanması sürecinde olduğu gibi deneme-yanılma yoluyla tespit edilmesi gerekmektedir.

6. SONUÇLAR (CONCLUSIONS)

Bu çalışmada, metin tabanlı dil tanıma için yeni bir yaklaşım önerilmiştir. Çalışmada, karakterlerin sırasal düzenlerine göre elde edilen desenlere göre dil tanıma işlemi gerçekleştirilmiştir. Dört farklı veri kümesi için elde edilen tanıma başarı oranları %86,20, %92,75, %100 ve %89,77 olarak gözlenmiştir. Önerilen yöntem ayrıca farklı uzunlukta metinler için denenmiş metin uzunluklarının büyük olması durumunda başarı oranının arttığı görülmüştür. Kabul edilebilir bir başarı oranı için metin uzunluklarının 500 byte ve üzeri uzunluklarda olması gerektiği gözlenmiştir. Önerilen yöntem ile elde edilen öznitelikler kullanılarak YSA, DVM, ÇTNB, FKA ve DS makine öğrenmesi yöntemler ile sınıflandırmıştır ve tüm yöntemler ile kabul edilebilir başarı oranları elde edilmiştir. Önerilen yaklaşımla elde edilen sonuçlar literatürde yayınlanan sonuçlar ve yaygın olarak kullanılan n-gram ile karşılaştırılmış ve önerilen yaklaşımın sonuçlarının başarılı olduğu görülmüştür. Önerilen yaklaşım DT'nın yanı sıra spam tanıma, metin kategorize etme gibi farklı metin madenciliği alanlarında kullanılabilme potansiyeline sahiptir.

KAYNAKLAR (REFERENCES)

1. Selamat A., Ng C.C., Arabic script web page language identifications using decision tree neural networks, Pattern Recognition, 44 (1), 133-144, 2011.
2. Takçı H., Ekinci E., Minimal feature set in language identification and finding suitable classification method with it, Procedia Technology, 1, 444–448, 2012.
3. Xafopoulos A., Kotropoulos C., Almpanidis G., Pitas I., Language identification in web documents using discrete HMMs, Pattern Recognition, 37 (3), 583-594, 2004.
4. Nie J.Y., Cross-language information retrieval, Synthesis Lectures on Human Language Technologies, 3 (1), 1-125, 2010.
5. Li H., Ma B., Lee C.H., A vector space modeling approach to spoken language identification, IEEE

Transactions on Audio, Speech, and Language Processing, 15 (1), 271-284, 2007.

6. Nakamura S., Markov K., Nakaiwa H., Kikui G., Kawai H., Jitsuhiro T., Zhang J.S., Yamamoto H., Sumita E., Yamamoto, S., The ATR multilingual speech-to-speech translation system, IEEE Transactions on Audio, Speech, and Language Processing, 14 (2), 365-376, 2006.

7. Kaya Y., Ertuğrul Ö.F., Tekin R., An Expert Spam Detection System Based on Extreme Learning Machine, Computer Science and Applications, 1 (2), 132-137, 2014.

8. Selamat A., Omatu S., Web page feature selection and classification using neural networks, Information Sciences, 158, 69-88, 2004.

9. Haltaş A., Alkan A., Karabulut M., Performance analysis of heuristic search algorithms in text classification, Journal of the Faculty of Engineering and Architecture of Gazi University, 30 (3), 417-427, 2015.

10. Gültepe Y., Ünalır M.O., Rendering ontology based Turkish national health data dictionary and enrichment with medical informatics standards, Journal of the Faculty of Engineering and Architecture of Gazi University, 29 (3), 637-644, 2014.

11. Mani I., Maybury M.T., Advances in automatic text summarization, 293, Cambridge: MIT press, Massachusetts-USA, 1999.

12. Chong L.K., Kamprath C.K., Machine translation and telecommunications system, U.S. Patent No 5497319, March 5, 1996.

13. Takcı H., Soğukpınar İ., Letter based text scoring method for language identification, International Conference on Advances in Information Systems, İzmir-Türkiye, 283-290, October 20-22, 2004.

14. Evans D.A., Grefenstette G.T., Tong X., Method of identifying the language of a textual passage using short word and/or n-gram comparisons, U.S. Patent No: US7359851, Washington, DC: U.S. Patent and Trademark Office, April 15, 2008.

15. Cavnar W.B., Trenkle J.M., N-gram-based text categorization, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas-Nevada-USA, 161–175, April 11-13, 1994.

16. Popescu M., Dinu L.P., Kernel methods and string kernels for authorship identification: The federalist papers case. International Conference on Recent Advances in Natural Language Processing (RANLP-07), Borovets-Bulgaria, September 27-29, 2007.

17. Popescu M., Grozea C., Kernel methods and string kernels for authorship analysis Notebook for PAN at CLEF, Conference and Labs of the Evaluation Forum, Rome-Italy, September 17-20, 2012.

18. Popescu M., Ionescu R.T., The Story of the Characters, the DNA and the Native Language, Eighth Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta-GA-USA, 270–278, June 13, 2013.

19. Ahmed B., Cha, S.H., Tappert C., Language identification from text using n-gram based cumulative frequency addition, Proceedings of Student/Faculty Research Day, CSIS, Pace University, 12.1-12.8, May 7, 2004.

20. Burçin K., Vasif N.V., Down syndrome recognition using local binary patterns and statistical evaluation of the system, Expert Systems with Applications, 38 (7), 8690-8695, 2011.

21. Takçı H., Güngör T., A high performance centroid-based classification approach for language identification, Pattern Recognition Letters, 33 (16), 2077-2084, 2012.

22. Prager J.M., Linguini: Language identification for multilingual documents, 32nd Annual Hawaii International Conference on Systems Sciences, Hawaii-USA, 1-11, January 5-8, 1999.

23. Suzuki I., Mikami Y., Ohsato A., Chubachi Y., A language and character set determination method based on N-gram statistics, ACM Transactions on Asian Language Information Processing, 1 (3), 269-278, 2002.

24. Ng C.C., Selamat A., Improved letter weighting feature selection on arabic script language identification, First Asian Conference on Intelligent Information and Database Systems, Dong Hoi City-Vietnam, 150-154, April 1-3, 2009.

25. Li Q., Chen Y.P., Personalized text snippet extraction using statistical language models, Pattern Recognition, 43 (1), 378-386, 2010.

26. Sibun P., Reynar J.C., Language identification: examining the issues, In: Proc.5th Symposium on Document Analysis and Information Retrieval, Las Vegas-Nevada-USA, 125–135, April 15-17, 1996.

27. Song Y., Dai L., Wang R., An automatic language identification method based on subspace analysis, IEEE International Conference on Multimedia and Expo, New York-NY-USA, 598-601, 28 Jun - 03 Jul 2009.

28. Takci H., Diagnosis of breast cancer by the help of centroid based classifiers, Journal of the Faculty of Engineering and Architecture of Gazi University, 31 (2), 323-330, 2016.

29. Özocak A., Yurtcu Ş., Prediction of compression index of fine-grained soils using statistical and artificial intelligence methods, Journal of the Faculty of Engineering and Architecture of Gazi University, 31 (3), 597-608, 2016.

30. Köklü M., Kahramanlı H., Allahverdi N., A new accurate and efficient approach to extract classification rules, Journal of the Faculty of Engineering and Architecture of Gazi University, 29 (3), 477-486, 2014.

31. Jiang C., Coenen F., Sanderson R., Zito M., Text classification using graph mining-based feature extraction, Knowledge-Based Systems, 23 (4), 302-308, 2010.

32. Tan S., An effective refinement strategy for KNN text classifier, Expert Systems with Applications, 30 (2), 290-298, 2006.

33. Botha G.R., Barnard E., Factors that affect the accuracy of text-based language identification, Computer Speech & Language, 26 (5), 307-320, 2012.

34. Hayta Ş.B., Takçı H., Eminli M., Language Identification Based on n-Gram Feature Extraction Method by Using Classifiers, IU-Journal of Electrical & Electronics Engineering, 13 (2), 1629-1639, 2013.

35. Yavanoğlu U., Sağıroğlu, Ş., Automatic web based language identification and translation system, Journal of the Faculty of Engineering and Architecture of Gazi University, 25 (3), 483-494, 2010.

36. Baldwin T., Lui M., Language Identification: The Long and the Short of the Matter, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles-USA, 229-237, June 1, 2010.

37. Kaya Y., Uyar M., Tekin R., Yıldırım S., 1D-local binary pattern based feature extraction for classification of epileptic EEG signals, Applied Mathematics and Computation, 243, 209-219, 2014.

38. Kaya Y., Ertuğrul Ö.F., A novel approach for spam email detection based on shifted binary patterns, Security and Communication Networks, 9 (10), 1216–1225, 2016.

39. Witten I.H., Frank E., Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann Publishers, San Francisco-USA, 2005.