KONUŞMA DİLİ KULLANILARAK DEMOGRAFİK BİLGİLERİN SINIFLANDIRILMASI

DEMOGRAPHIC INFORMATION CLASSIFICATION EXPLOITING SPOKEN LANGUAGE

H. İrem Türkmen, Banu Diri, Göksel Biricik, Reşit Doğan

Bilgisayar Mühendisliği Bölümü
Yıldız Teknik Üniversitesi
{irem,banu,goksel}@ ce.yildiz.edu.tr, resit@yildiz.edu.tr

ÖZETÇE

Kişilerin yüz ve ses özelliklerinden faydalanılarak yaş, cinsiyet, ırk gibi demografik bilgilerine ulaşabilmek için yapılan çalışmalar son yıllarda hız kazanmıştır. Bu çalışmada, doğal dil ile yazılmış dokümanlardan demografik bilgileri tahmin eden bir sistem geliştirilmiş, böylece görüntü ve ses özelliklerinin yanı sıra konuşma dili özelliklerinin de kişilerin demografik profilleri ile ilişkili olduğu gösterilmiştir. İlk olarak, konuşma dilinde yazılmış dokümanlardan farklı sayıda özellik vektörleri çıkarılmış, daha sonra geliştirmiş olduğumuz özellik azaltma yöntemi ve Korelasyon Tabanlı Özellik Seçici algoritmaları kullanılarak vektörlerin boyutları azaltılmıştır. Son olarak da Naïve Bayes, Destek Vektör Makineleri ve K-En Yakın Komşuluk algoritmaları ile sınıflandırma başarıları değerlendirilmiştir.

ABSTRACT

Recently, extracting the demographic information like age, gender and race by using speech and face attributes takes much attention in the literature. In this research, we have focused on the implementation of a demographic information classification system and proved the relationship between spoken language and demographic profile of people. In the first step, the feature vectors of spoken language were extracted then dimensions of the feature vectors were reduced by our feature reduction method and Correlation Based Feature Selection method. Finally, the success of Naïve Bayes, Support Vector Machine and K-Nearest Neigbour classification algorithms was evaluated.

1. GİRİŞ

Doğal dil işlemenin bir alt dalı olan doküman sınıflandırma, web sayfalarının hiyerarşik olarak düzenlenmesi, bir dokümanın türünün bulunması, dokümanın yazarının tahmin edilmesi, yazarın cinsiyetinin tespit edilmesi gibi birçok uygulama alanına sahiptir. Sınıflandırmadaki amaç, farklı yaklaşımlar kullanarak bir dokümanın belirleyici özelliklerini çıkarmak ve bu özelliklerin yardımıyla önceden belirlenmiş belli sayıdaki kategorilerden hangisine daha yakın olacağını belirlemektir. Son yıllarda yüz tanıma çalışmaları içerisinde yaş, cinsiyet ve ırk gibi demografik bilgilere göre sınıflandırma çalışmaları yapılmaktadır [1]. Bu çalışmada ise, sınıf etiketi olarak belirlediğimiz yaş, medeni durum, eğitim durumu ve cinsiyet gibi demografik bilgileri bilinen 484 kişiye "Yaşadığınız bir ahlaki çatışmaya örnek verir misiniz?" sorusu sorulmuştur. Bu soruya verilen cevaplar içerisinden belli özellikler çıkarılarak, farklı özellik vektörleri elde edilmiş ve eğiticili öğrenme metotları kullanılarak eğitilen sistemin, demografik bilgileri bilinmeyen kişileri yaş, cinsiyet, medeni hal ve eğitim durumu olmak üzere dört ayrı kategoride sınıflandırması sağlanmıştır. Böylece kişinin düşüncelerini ifade ettiği konuşma dili ile yazılmış metinlerden yararlanılarak demografik bilgilerin tahmin edilebileceği gösterilmiştir.

Doküman sınıflandırma sistemlerinin ilk örnekleri 70'li yıllarda otomatik doküman indeksleme olarak karşımıza çıkmıştır. Belirli bir konu için özel sözlükler oluşturulmuş ve sözlük içerisindeki kelimeler birer kategori gibi algılanarak dokümanlar sınıflandırılmıştır. Mosteller ve Wallace, yazarlık özelliklerini çıkararak Bayesian analizi ile yazar tanıma yapmışlardır [2]. Burrows [3] en fazla sıklıkta kullanılan kelimeleri, Brinegar [4] kelimelerin uzunluğunu, Morton [5] cümlelerin uzunluğunu, Brainerd [6] ortalama hece sayısını, Holmes [7] kullanılan ayrık kelime sayısını ve dokümanın uzunluğunu, Twedie ve Baayen [8] ise farklı kelime sayısının toplam kelime sayısına oranını kullanarak doküman sınıflandırma yapmışlardır. Stamatatos ve arkadaşları [9] doğal dil işleme hazır paket programını kullanarak bir dizi stil belirleyici (style marker) elde etmiş ve bunlardan yararlanarak yazar tanıma yapmışlardır. Fürnkranz [10] n-gram özelliklerini, Tan ve arkadaşları [11] 2-gram'ları kullanarak doküman sınıflandırmada performansı arttırmışlardır. Diri ve Amasyalı [12] bir dokümanın yazarını ve türünü belirlemede kullanılmak üzere 22 adet stil belirleyicisi oluşturmuş ve bunları kullanan bir sınıflandırma sistemi geliştirmişlerdir. Yine aynı yazarlar [13] 2 ve 3-gram'ları kullanarak dokümanın türünü, yazarını ve yazarının cinsiyetini belirleme üzerine çalışmışlar ve bir başka çalışmalarında ise [14] Türkçe'nin biçimbilim yapısından yararlanılarak çıkarılmış farklı özellik vektörleri ile yazar tanıma yapmış ve sonuçları karşılaştırmalı olarak sunmuşlardır.

978-1-4577-0463-511/11/$26.00 ©2011 IEEE

Makalenin ikinci bölümünde veri setinden, üçüncü bölümde kullanılan yöntemlerden ve dördüncü bölümde deneysel sonuçlardan bahsedilmektedir.

2. VERİ SETİ VE ÖZELLİK ÇIKARTMA

Bu çalışmada kullanılan veri seti, İstanbul Üniversitesi-Edebiyat Fakültesi, Psikoloji Bölümü tarafından "Yaşanan Ahlaki Çatışma Örnekleri üzerinden Türkiye'de Gündelik Ahlak Anlayışı" isimli çalışmada kullanılmak üzere hazırlanmıştır. Veri seti oluşturulurken 484 kişiye "Yaşadığınız bir ahlaki çatışmaya örnek verir misiniz?" sorusu sorulmuş ve soruya konuşma dili ile cevap vermeleri istenmiştir. Alınan cevaplar, soruya cevap veren kişilerin demografik bilgileri ile birlikte kaydedilmiştir. Tablo 1'de veri setinde bulunan cevaplardan iki örnek verilmiştir.

Tablo 1: Örnek Cevaplar

Ahlaki çatışma yaşamam için belirli dogmatik ahlak kurallarının altında yaşamam gerekirdi. Fakat böyle dayatmaları kabul eden biri değilim. Dolayısıyla kendi ahlak anlayışımın dışında bir hareket yapmadım
Ortaokulda bilgisayar laboratuvarındayken bilgisayarım açılmadı, hocama söyledim ama hocam ilgilenmedi. Hocam ilgilenmediği için çok sinirim bozuldu, o sinir bozukluğuyla birlikte bilgisayarın voltaj ayarını değiştirdim ve en yakın arkadaşımdan pc yi açmasını istedim ve arkadaşım bilgisayarı açtığında dumanlar çıkmaya başladı yani pc bozuldu. Ve şu an bile çok yakın arkadaşıma suçu yıkmak ve tüm parayı ödettirmiş olmak çok ağrıma gidiyor. Bu olaydan korktuğum için, daha doğrusu hain olmaktan korktuğum için olayı arkadaşıma hala anlatamıyorum kendimce çözümü bu.

Sorulara cevap veren kişilerin demografik bilgileri ve bu bilgilere göre dağılımları şu şekildedir; Cinsiyet (Kadın:257, Erkek:227), Yaş (10-20 yaş arası:147 (Yaş Grup 1), 21-35 yaş arası:223 (Yaş Grup 2), 36-66 yaş arası:103 (Yaş Grup 3), bilinmeyen:11), Medeni durum (evli:143, bekar:327, bilinmeyen:14), Eğitim derecesi (hiç okumamış - ortaokul mezunu:140 (Derece 1), lise-üniversite öğrencisi:215 (Derece 2), üniversite-yüksek lisans/doktora mezunu:129 (Derece 3)).

Yapılan çalışmada veri seti içerisinde yer alan her kelimenin kök ve tipinin bulunması için açık kaynak kodlu doğal dil işleme aracı Zemberek [15], 2-gramlar içinse YTU-Kemik grubu tarafından geliştirilen text2arff (v3.0) [16] aracı kullanılmıştır. Daha sonra yine text2arff aracı kullanılarak, kelime köklerinin ters doküman frekanslarından (idft) oluşturulmuş ve kök, tip, 2-gram'larının ters doküman frekanslarının birleşiminden oluşturulmuş arff formatındaki özellik vektörleri elde edilmiştir. Elde edilen iki farklı tipteki özellik vektörleri, özellik azaltma ve sınıflandırma algoritmalarına ayrı ayrı verilmiş ve böylece farklı özellik çıkarma metotlarının sınıflandırma başarısını nasıl etkilediği incelenmiştir.

3. ÖZELLİK AZALTMA VE SINIFLANDIRMA

Kelimelerin köklerinin ters doküman frekanslarından ve kök, tip, 2-gram'ların ters doküman frekanslarının birleşiminden oluşturulmuş olan özellik vektörlerinin boyutları çok büyük olduğundan, daha yüksek sınıflandırma başarısı alabilmek için, özellik azaltmaya gidilmiştir. Özellik azaltımında Weka içerisinde yer alan Korelasyon-tabanlı Özellik Seçici (KÖS) / Correlation-based Feature Selection (CFS) ile YTU Özellik Azaltıcı yöntemleri tercih edilmiştir. [17,18,19].

Korelasyon-tabanlı Özellik Seçici (KÖS), diğer özelliklerle düşük korelasyonlu, sınıf değişkeni ile yüksek korelasyonlu olan özellikleri seçen bir yöntemdir.

YTU Özellik Azaltıcı, dokümanlardaki terim olasılıklarının sınıflara olan izdüşümü alınıp, bu olasılıklar toplanarak her terimin sınıfları ne kadar etkilediğini bulan bir yöntemdir. Böylece özellik sayısı sınıf sayısına indirgenmiş olmaktadır. n adet doküman ve k adet sınıf olduğu varsayılarak;

$n_{i,j}$:t_i teriminin, s_j dokümanında kaç kere geçtiğini

n_i :t_i terimine sahip olan doküman sayısını vermiş olsun. Bir t_i teriminin, c_k sınıfında kaç kere geçtiği (1) ile hesaplanabilir. Bir s_j dokümanında yer alan t_i teriminin, c_k sınıfını ne kadar etkilediği de (2) ile bulunabilir.

$$nc_{i,k} = \sum_{j} n_{i,j}, \quad s_j \in c_k \tag{1}$$

$$w_{i,k} = \log\left(nc_{i,k}+1\right) \times \log\left(\frac{n}{n_i}\right) \tag{2}$$

Bir dokümandaki tüm terimlerin c_k sınıfına olan toplam etkisi (3) ile hesaplanır. Sonuçta, i adet terim sınıf sayısı k kadar boyutlu bir hiper düzleme izdüşürülür. Bu işlem tüm dokümanlar için uygulandığında n satır ve k sütundan oluşan indirgenmiş sonuç matrisi ve (4) ile de normalize edilmiş yeni özellikler elde edilir.

$$Y_{j,k} = \sum_{i} w_{i,k}, \quad w_i \in s_j \tag{3}$$

$$T_k = \frac{Y_{j,k}}{\sum_{k} Y_{j,k}} \tag{4}$$

Dört ayrı kategoride yapılan sınıflandırmada kullanılan ham özellik vektörlerinin ve bu özellik vektörlerinin KÖS ve YTU algoritmaları ile indirgenmiş hallerinin boyutları Tablo 2'de verilmektedir. YTU yöntemi, özellik azaltmada doküman terim olasılıklarını kullandığı için, kök, tip ve 2-gram'ların frekanslarının birleşimlerinden oluşturulmuş özellik vektörlerinde kullanılmamıştır.

Tablo 2: Özellik vektörlerinin boyutları

		Kök	**Kök+Tip +2-gram**
Cinsiyet (2 Sınıf)	Özellik	2964	5209
	KÖS	70	86
	YTU	2	-
Eğitim (3 Sınıf)	Özellik	2964	4676
	KÖS	50	87
	YTU	3	3
Yaş (3 Sınıf)	Özellik	2945	3783
	KÖS	7	64
	YTU	3	-
Medeni Durum (2 Sınıf)	Özellik	2929	4631
	KÖS	57	75
	YTU	2	-

Demografik bilgilere göre sınıflandırma yaparken eğiticili sınıflandırma yöntemlerinden Naïve Bayes (NB), Destek Vektör Makinesi (DVM)/Support Vector Machine (SVM) ve K-En Yakın Komşuluk (KEK)/K-Nearest Neigbour (KNN) algoritmaları kullanılarak başarıları değerlendirilmiştir. Adı geçen bütün sınıflandırma algoritmaları WEKA[20] aracı ile gerçekleştirilmiştir.

Naïve Bayes (NB) Klasik Naïve Bayes algoritması genelde kelimelerin ve sınıfların birleşik olasılıkları ile bir dokümanın sınıfının belirlenmesinde kullanılır. Bu çalışmada ise özellikler kelimelerin frekansları değildir ve sürekli dağılımlara sahip olduklarından klasik Naïve Bayes yerine George'un [21] çalışmasında önerilen Naïve Bayes versiyonu kullanılmıştır.

Destek Vektör Makinesi (DVM) Sınıfları birbirinden ayıran marjini en büyük, doğrusal bir ayırt edici fonksiyon bulunmasını amaçlar. Doğrusal olarak ayrılamayan örnekler için veriler, doğrusal olarak ayrılabildikleri daha yüksek boyutlu başka bir uzaya taşınır ve sınıflandırma o uzayda yapılır.

K-En Yakın Komşuluk (KEK) Sınıfı bulunacak olan verinin özellik vektörü değerine en yakın olan *k* adet özellik vektörlerinin bulunması prensibine dayanır. En yakın özellik vektörleri bulunurken Euclidean mesafesi kullanılmıştır. Bulunan *k* adet vektör en fazla hangi sınıfa ait ise, o sınıfı etiketi sınıflandırılacak olan verinin sınıfı olarak belirlenir.

4. DENEYSEL SONUÇLAR

Bu bölümde farklı özellik azaltma ve sınıflandırma algoritmaları kullanılarak elde edilen demografik bilgilere göre sınıflandırma başarıları incelenmiştir. Verilen sonuçlar, DVM ile sınıflandırmada tüm kategorilerde en iyi başarı oranını veren polinomal çekirdek, KEK ile sınıflandırmada ise k=5 değeri kullanılarak elde edilmiştir. Tüm sınıflandırma problemlerinde sistem başarısı, 10 katlı çapraz geçerlilik testi kullanılarak değerlendirilmiştir. Tablo 3'de cinsiyet tanıma için elde edilen ortalama başarı oranları, Tablo 4'de ise medeni hal tanıma için elde edilen ortalama başarı oranları görülmektedir.

Tablo 3: Cinsiyet Tanıma Başarıları (%)

	Kök	KÖS(Kök)	YTÜ(Kök)	Kök,Tip, 2-gram	KÖS(Kök, Tip,2-gram)
NB	55.8	66.5	77.1	54.5	67.6
KEK	53.1	65.5	**78.1**	52.1	68.6
DVM	58.1	53.1	74.8	58.1	72.3

Tablo 4: Medeni Hal Sınıflandırma Başarıları (%)

	Kök	KÖS(Kök)	YTU(Kök)	Kök,Tip, 2-gram	KÖS(Kök, Tip,2-gram)
NB	69.6	82.0	67.0	69.4	81.7
KEK	69.0	77.4	76.4	69.4	79.6
DVM	71.7	**86.1**	69.6	71.9	83.8

Tablo 3' de görüldüğü gibi, en yüksek cinsiyet tanıma başarı oranı YTU yöntemi ile azaltılan özelliklerin KEK yöntemi ile sınıflandırılması sonucu elde edilmiştir. Medeni hale göre sınıflandırmada ise Korelasyon Tabanlı Özellik Seçici ve DVM yöntemleri en yüksek başarı oranlarını vermiştir.

Bu yöntemler ile cinsiyet ve medeni hal tanımaya ait hata matrisleri Tablo 5 ve Tablo 6'da görülmektedir.

Tablo 5: Cinsiyet Tanıma Hata Matrisleri

		Hesaplanan		
		erkek	**kadın**	çağrı[1]
Gerçek	**erkek**	164	63	0.72
	kadın	43	214	0.83
	kesinlik[1]	0.79	0.77	

Tablo 6: Medeni Hal Tanıma Hata Matrisleri

		Hesaplanan		
		bekar	**evli**	*çağrı*
Gerçek	**bekar**	313	14	0.96
	evli	52	91	0.64
	kesinlik	0.86	0.87	

Cinsiyet tanımada erkek sınıfı için F-ölçüm[1] oranı 0.76, kadın sınıfı için ise 0.80 olarak, medeni hal tanımada bekar sınıfı için F-ölçüm oranı 0.91, evli sınıfı için ise 0.73 olarak hesaplanmıştır.

Eğitim derecesine göre sınıflandırma başarı oranları Tablo 7'de gösterilmektedir.

Tablo 7: Eğitim Derecesi Sınıflandırma Başarıları (%)

	Kök	KÖS(Kök)	YTU(Kök)	Kök,Tip, 2-gram	KÖS(Kök, Tip,2-gram)
NB	45.0	52.5	**95.2**	44.4	56
KEK	41.3	47.5	94.8	41.3	46.1
DVM	45.2	56.2	94.6	49.6	56.2

Tablo 7'de görüldüğü gibi, en yüksek eğitim durumuna göre sınıflandırma başarı oranı YTU yöntemi ile azaltılan özelliklerin Naïve Bayes yöntemi ile sınıflandırılması sonucu elde edilmiştir. Bu yöntemler ile eğitim derecesine göre sınıflandırma hata matrisi Tablo 8'de görülmektedir.

Tablo 8: Eğitim Derecesi Sınıflandırma Hata Matrisi

		Hesaplanan			
		Derece 1	**Derece 2**	**Derece 3**	*çağrı*
Gerçek	**Derece 1**	130	8	2	0.93
	Derece 2	7	207	1	0.96
	Derece 3	5	0	124	0.96
	kesinlik	0.92	0.96	0.98	

Eğitim derecesine göre sınıflandırmada, Derece 1 için F-ölçüm oranı 0.92, Derece 2 için 0.96, Derece 3 için ise 0.97 olarak hesaplanmıştır.

Yaşa göre sınıflandırma başarı oranları Tablo 9'da görülmektedir.

Tablo 9: Yaş için Sınıflandırma Başarıları (%)

	Kök	KÖS(Kök)	YTU(Kök)	Kök,Tip, 2-gram	KÖS(Kök, Tip,2-gram)
NB	50.1	57.7	84.6	48.6	56.65
KEK	33.2	55.0	86.3	39.9	55.39
DVM	50.7	58.8	**87.3**	52.43	66.59

[1] Çağrı (recall), kesinlik (precision), F-ölçüm (f-measure)

Tablo 9'da görüldüğü gibi, en yüksek yaşa göre sınıflandırma başarı oranı YTU yöntemi ile azaltılan özelliklerin, DVM yöntemi ile sınıflandırılması sonucu elde edilmiştir. Bu yöntemler ile yaşa göre sınıflandırma hata matrisi Tablo 10' da görülmektedir.

Tablo 10: Yaş Grubuna Göre Hata Matrisi

		Hesaplanan			
		YaşGrup1	**YaşGrup2**	**YaşGrup3**	*çağrı*
Gerçek	**YaşGrup1**	114	33	0	0.78
	YaşGrup2	10	204	9	0.92
	YaşGrup3	8	0	95	0.92
	kesinlik	0.86	0.86	0.91	

Yaş gruplarına göre sınıflandırmada, YaşGrup 1 için F-ölçüm oranı 0.82, YaşGrup 2 için 0.89, YaşGrup 3 için ise 0.92 olarak bulunmuştur.

Sonuçlar incelendiğinde, demografik bilgilere göre sınıflandırma probleminde, özellik azaltma metotlarının kullanımının sınıflandırma başarılarını çok fazla etkilediği görülmektedir. Kelime kökü tabanlı özellik çıkarma metotları dört tip sınıflandırma probleminde de en iyi sonuçları vermiştir.

5. SONUÇLAR

Bu çalışmada konuşma dilinden yararlanılarak oluşturulan dokümanlar kullanılarak demografik bilgilerin tahmin edilmesi için bir sistem tasarlanmış ve gerçeklenmiştir. Bunun için kelimelerin kökleri, tipleri ve 2-gram'ları kullanılarak çıkarılan özelliklerin, KÖS ve YTU yöntemleri kullanılarak özellik sayıları azaltıldıktan sonra, farklı sınıflandırma algoritmaları ile sınıflandırılma başarıları değerlendirilmiştir. Bu sayede kişilerin yaş, cinsiyet, eğitim durumu, medeni hal gibi bilgilerinin konuşma dili ile ilişkili olduğu gösterilmiştir. Çalışma, elde edilen yüksek başarı oranları sayesinde psikoloji, sosyoloji, adli bilimler gibi bilim dallarında kullanılma imkânı bulabilecek ve ileriki çalışmalara ışık tutacak niteliktedir.

6. TEŞEKKÜR

Bu çalışmada kullanılan veri setini bizimle paylaşan İstanbul Üniversitesi, Edebiyat Fakültesi, Psikoloji Bölümü, Sosyal Psikoloji A.B.D.'ında görev yapmakta olan Yrd.Doç.Dr. Sevim Cesur ve Yrd.Doç.Dr. Göklem Tekdemir'e teşekkür ederiz.

7. KAYNAKÇA

[1] Z. Yang ve H. Ai, "Demographic Classification with Local Binary Patterns", *The 2nd Int.Conf.on Biometrics*, Italy, 2007.

[2] F. Mosteller ve D.L. Wallace, "Applied Bayesian and Classical Inference: The Case of the Federalist Papers", *Reading, MA:Addison-Wesley*, 1984.

[3] J.F. Burrows, "Not unless you ask nicely: the interpretative nexus between analysis and information", *Literary Linguist Comput*, 1992, 7:91-109.

[4] C.S. Brinegar, "Mark Twain and the Quintus Curtius Snodgrass letters: A statistical test of authorship", *Journal of the American Statistical Association*, 1963, 58:85-96.

[5] A.Q. Morton, "The Authorship of Greek Prose", *Journal of the Royal Statistical Society, Series A*, 1965, 128:169-233.

[6] B. Brainerd, Weighting Evidence in Language and Literature: A Statistical Approach. *University of Toronto Press*, 1974.

[7] D.I. Holmes, "Authorship Attribution", *Compute Humanities*, 1994, 28:87-106.

[8] F. Tweedie ve H. Baayen, "How Variable may a Constant be Measures of Lexical Richness in Perspective", *Computers and the Humanities*, 1998, 32(5), 23-352.

[9] E. Stamatatos, N. Fakotakis and G. Kokkinakis, Automatic Text Categorization in Terms of Genre and Author, *Computational Linguistics*, 2000, pp.471-495.

[10] J. Fürnkranz, "A Study using n-gram Features for Text Categorization", *Austrian Research Institute for Artifical Intelligence*, 1998.

[11] C.M. Tan, Y.F. Wang ve C.D. Lee, The Use of Bi-grams to Enhance, *Journal of Information Processing and Management*, 2002, Vol:30 No:4 pp.529-546.

[12] B. Diri ve M.F. Amasyalı, "Automatic Author Detection for Turkish Texts", *Artificial Neural Networks and Neural Information Processing*, 138-141, 2003.

[13] M.F. Amasyalı and B.Diri, "Automatic Written Turkish Text Categorization in Terms of Author, Genre and Gender", *11th International Conference on Applications of Natural Language to Information Systems*, Austria, 2006.

[14] M.F. Amasyalı, B. Diri ve F. Türkoğlu, "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", *15th Turkish Symposium on Artificial Intelligence and Neural Networks*, Muğla, 2006.

[15] http://code.google.com/p/zemberek/

[16] http://www.kemik.yildiz.edu.tr/?id=29

[17] H.K.Yıldız, M. Gençtav, N. Usta, B. Diri ve M.F. Amasyalı, "Metin Sınıflandırmada Yeni Özellik Çıkarımı", *IEEE-15. Sinyal İşleme, İletişim ve Uygulamaları Kurultayı*, Eskişehir, 2007.

[18] Z. Kaban ve B. Diri, "Yapay Bağışıklık Sistemleri ile Türkçe Metinlerde Tür ve Yazar Tanıma", *IEEE-16. Sinyal İşleme, İletişim ve Uygulamaları Kurultayı*, Aydın, 2007.

[19] G. Biricik, B. Diri, "Impact of a New Attribute Extraction Algorithm on Web PageClassification", *5th International Conference on Data Mining*, USA, 2009.

[20] www.cs.waikato.ac.nz/ml/weka/

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