Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi

M. Fatih Amasyalı[1], Banu Diri[1], Filiz Türkoğlu[2]

Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği
34349 İstanbul-Türkiye

[1]{mfatih,banu}@ce.yildiz.edu.tr, [2]filizturkoglu@gmail.com

Özet. Bu çalışmanın amacı, yazarı bilinmeyen bir dokümanın önceden belirlenmiş ve yazarlık özellikleri çıkarılmış 18 farklı yazardan hangisine ait olduğunu tespit etmektir. Yazarı bulunması istenen dokümanlardan istatistiksel veriler, kelime sözlüğünün zenginliği, sık geçen Türkçe kelimeler, dilbilgisi özellikleri, n-gram'lar ve bunların çeşitli birleşimleri kullanılarak on farklı özellik vektörü elde edilmiştir. Daha sonra Naïve Bayes, SVM, C 4.5 ve RF sınıflandırma yöntemleri kullanılarak her bir özellik vektörünün sınıflandırmadaki başarıları karşılaştırılmıştır. En başarılı sınıflandırma yöntemi Naïve Bayes ve SVM'dir. Dokümanların sınıflandırılmasında, n-gram'ların kullanılmasının yazarlık özelliklerine göre daha başarılı olduğu gözlemlenmiştir. Bununla birlikte n-gram ve yazarlık özellikleri birlikte kullanıldıklarında daha başarılı sonuçlar elde edilmiştir.

Abstract. The goal of this study is classifying documents according to their author. Ten different feature vector is constructed from token-level features, vocabulary richness, grammatical statistics, frequencies of function words and n-gram frequencies. Naïve Bayes, Support Vector Machine, C 4.5 and Random Forest are used as classification methods and the results are given comparatively. The most succsessfull classifiers are Naïve Bayes and SVM. The n-gram models are more successfull than other features. However, the most accurate results are obtained when all faetures are used

1 Giriş

Doküman sınıflandırmadaki amaç, bir dokümanın özelliklerine bakılarak önceden belirlenmiş belli sayıdaki kategorilerden hangisine dahil olacağını belirlemektir. Doküman sınıflandırma bilgi alma (information retrieval), bilgi çıkarma (information extraction), doküman indeksleme, doküman filtreleme, otomatik olarak metadata elde etme ve web sayfalarını hiyerarşik olarak düzenleme gibi pek çok alanda önemli bir rol oynamaktadır. Doküman sınıflandırmadaki problemlerden biri, kime ait olduğu bilinmeyen veya yazarının kimliğinden şüphelenilen dokümanların yazarının tahmin

edilmesi, bir diğer problem de dokümanın türünün veya yazarının cinsiyetinin belirlenmesidir.

Doküman sınıflandırma sistemlerinin ilk örnekleri 70'li yıllarda karşımıza otomatik doküman indeksleme olarak çıkmıştır. Belirli bir konu için özel sözlükler oluşturulmuş ve bu sözlük içerisindeki kelimeler birer kategori gibi algılanarak dokümanlar sınıflandırılmıştır. Mosteller ve Wallace çalışmalarında yazarlık özelliklerini belirlemişler ve Bayesian analizi kullanmışlardır [1]. Burrows [2] çalışmasında en fazla sıklıkta kullanılan kelimeleri, Brinegar [3] kelimelerin uzunluğunu, Morton [4] cümlelerin uzunluğunu, Brainerd [5] ortalama hece sayısını, Holmes [6] kullanılan kelime sayısını ve dokümanın uzunluğunu, Twedie ve Baayen [7] farklı kelime sayısının toplam kelime sayısına oranını kullanmışlardır. Stamatatos ve arkadaşları [8] doğal dil işleme hazır paket programını kullanarak bir dizi stil belirleyici (style marker) elde etmiş ve bunlardan yararlanarak yazar tanıma yapmışlardır. Fürnkranz [9] n-gram (2 ve 3 uzunluğunda) özelliklerini, Tan ve arkadaşları [10] 2-gram'ları (bi-gram) kullanarak bir algoritma geliştirmiş ve doküman sınıflandırmada performansı arttırmışlardır. Çatal ve arkadaşları [11], n-gram'ları kullanarak NECL adını verdikleri bir sistem geliştirmişlerdir. 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 Amasyalı ve Diri [13] 2 ve 3-gram'ları kullanarak dokümanın yazarını, türünü ve yazarının cinsiyetini belirme üzerine çalışmışlardır.

Bu çalışmada Türkçe dokümanlardan oluşan bir derlem kullanılmıştır. Türk dilinin 2 ve 3-gram'ları, sık geçen Türkçe kelimeler (stop words), dilbilgisel ve istatistiksel özellikler kullanılarak on farklı özellik vektörü çıkarılmıştır. Daha sonra makine öğrenmesi yöntemlerinden olan Naïve Bayes, Destek Vektör Makinesi (Support Vector Machine), C 4.5 ve Random Forest kullanılarak bir dokümanın yazarının belirlenmesi gerçekleştirilmiştir. Makalenin ikinci bölümünde yazarlık özelliklerinden, üçüncü bölümde ise sınıflandırma algoritmalarından bahsedilmiştir. Dördüncü ve beşinci bölümlerde ise sırasıyla deneysel sonuçlar ve sonuç bölümü yer almaktadır.

2 Yazarlık Özellikleri

Sürekli okuduğumuz bir günlük gazetenin köşe yazılarını bir müddet sonra yazarlarının isimleri yayınlanmadan okuduğumuzda, o yazının kime ait olduğunu rahatlıkla söyleyebiliriz. Kısacası her yazarın kendine has bir üslubu vardır. Yazarların kendilerine has olan bu tarz yazarlık özelliği (authorship attribution) olarak adlandırılır.

2.1 Derlem (corpus)

Bu çalışmada kullanılacak olan derlem için günlük gazetelerimizden olan Hürriyet (www.hurriyet.com.tr), Vatan (www.vatanim.com.tr) ve Sabah (www.sabah.com.tr)' dan politika, güncel, spor ve magazin gibi konularda yazılar indirilmiştir. Elimizdeki

mevcut derlem 18 farklı yazarın her birine ait 35 farklı yazısından oluşan, 630 adet dokümandan meydana gelmektedir.

2.2 N-gram

n-gram, *n* karakterden oluşan bir dizidir. N değeri teoride çok büyük bir sayı olabilir ancak uygulamada *n* değeri 4 veya 5 olarak seçilmelidir. Bu çalışmada en büyük *n* değeri üç olarak kabul edilmiş ve mevcut derlemden yararlanarak önce bi-gram (n=2) daha sonra da tri-gram'lar (n=3) bulunmuştur. "yazar" kelimesinin n-gram'larını bulacak olursak:
bi-grams: ya, az, za, ar, r_
tri-grams: yaz, aza, zar, ar_

2.3 Özellik Vektörleri

Beş farklı özellik vektörü çıkarılmış ve her bir özellik vektörüne özellik seçimi uygulanarak 5 tane daha özellik vektörü elde edilmiştir. Bu 10 özellik vektörünün sınıflandırma başarısı araştırılmıştır. Daha önceki bölümde bahsedildiği gibi amacımız yazarın sahip olduğu yazım stilini belirlemektir. Bu bölümde bahsedilen özellik azaltıcı olarak WEKA (www.cs.waikato.ac.nz/ml/weka/) paketi içerisinde yer alan CfsSubsetEval fonksiyonu kullanılmıştır.

Birinci özellik vektörü (Vg)
Birinci özellik vektörü 641 özellikten oluşmaktadır. Bu özellik vektöründen *Vg* olarak bahsedilecektir. İçerisinde istatistiksel verileri, kelime sözlüğünün zenginliğini, sık geçen Türkçe kelimeleri ve dilbilgisi özelliklerini bulunduran bir vektördür.

İstatistiksel özellikler sırasıyla, doküman içerisinde yer alan kelime sayısı, ortalama cümle uzunluğu, cümle sayısı, virgül-nokta-iki nokta üst üste-noktalı virgül-üç nokta yan yana sayısı, ünlem sayısı ve soru işareti sayısı olup 10 adettir.

Kelime sözlüğü zenginliği ise yazarın kullandığı kelime zenginliğini gösterir. Bu grup üç özelliğe sahiptir ve bunlar sırasıyla, hapax legomena, hapax dislegomena ve type/token oranıdır. Hapax legomena, doküman içerisinde kelimenin bir kez kullanılmış, hapax dislegomena ise kelimenin sadece iki kez kullanılmış olmasıdır. Tüm kelime sayısının hapax legomena ve dislegomena sayısına oranı yazar belirmede önemli bir ayıraçtır. Type/token oranı ise dokümanın sahip olduğu sözlüğün doküman içerisindeki toplam kelime sayısına oranıdır.

Sıkça geçen Türkçe kelimeler (stop word veya funcion word) ise sözcüksel olarak anlamı az veya belirsiz olan sözcük gruplarıdır. İlgeç, adıl, bağlaçlar, yardımcı fiiller bu gruba girmektedir. *Neden, ayrıca, böylece, adeta, daima, belki* örnek olarak verilebilir. Bu grupta toplam 620 farklı özellik bulunmaktadır.

Dilbilgisi özellikleri grubunda toplam 8 tane özellik bulunmaktadır. Bunlar sırasıyla cümle içerinde yar alan isim, fiil, sıfat, zarf, zamir, bağlaç, ünlem ve edattır.

İkinci özellik vektörü (Vga)
Birinci özellik vektörümüz olan *Vg*'nin 641 adet özelliği, özellik azaltıcı CfsSubsetEval fonksiyonu kullanılarak 24 özelliğe düşürülmüştür. Özellik sayısı azaltıldığında sınıflandırma işleminin süresi kısalmaktadır. 24 özelliğe sahip bu vektörü çalışmamızda *Vga* olarak kullanacağız.

Üçüncü ve dördüncü özellik vektörleri (*Vbg – Vbga*)
Üçüncü özellik vektörünü, derlemi kullanarak elde ettiğimiz 2-gram'lar oluşturmaktadır. 2-gram'ların sayısı çok fazla olduğundan belli bir frekansın üzerinde (bu çalışmada eşik değeri 75 olarak seçilmiştir) yer alan 2-gram'lar seçilmiştir. Bu vektörün özellik sayısı 470 olup, *Vbg* olarak adlandırılmıştır. Dördüncü özellik vektörü ise özellik azaltıcı CfsSubsetEval fonksiyonu kullanılarak, *Vbg* vektörünün özellik sayısının 25'e düşürülmesiyle elde edilmiştir. Bu yeni vektör *Vbga* olarak kullanılacaktır.

Beşinci ve altıncı özellik vektörleri (*Vtg – Vtga*)
Beşinci özellik vektörünü, derlemi kullanarak elde ettiğimiz 3-gram'lar oluşturmaktadır. 3-gram'ların sayısı çok fazla olduğundan belli bir frekansın üzerinde (bu çalışmada eşik değeri 75 olarak seçilmiştir) yer alan 3-gram'lar seçilmiştir. Bu vektörün özellik sayısı 1037 olup, *Vtg* olarak adlandırılmıştır. Altıncı özellik vektörü ise özellik azaltıcı CfsSubsetEval fonksiyonu kullanılarak, *Vtg* vektörünün özellik sayısının 60'a düşürülmesiyle elde edilmiş ve *Vtga* olarak adlandırılmıştır.

Yedinci ve sekizinci özellik vektörleri (*Vbtg – Vbtga*)
Yedinci özellik vektörü *Vbg* ve *Vtg* vektörlerinin birleşiminden oluşmuştur. Bu özellik vektörü içerisinde 1507 özellik bulunup *Vbtg* olarak adlandırılmıştır. Yine aynı sebepten yedinci özellik vektörü, özellik azaltıcı CfsSubsetEval fonksiyonunu kullanarak *Vbtg* vektörünün özellik sayısının 61'e düşürülmesiyle elde edilmiştir. Bu yeni vektöre *Vbtga* adı verilmiştir.

Dokuzuncu ve onuncu özellik vektörleri (*Vgbtg – Vgbtga*)
Dokuzuncu özellik vektörü *Vg* ile *Vbtg* vektörlerinin birleşiminden oluşup toplam 2148 özellik içermektedir. Bu özellik vektörü *Vgbtg* olarak kullanılacaktır. *Vgbtg* vektörünün özellik sayısı, özellik azaltıcı CfsSubsetEval fonksiyonu kullanılarak 69 özelliğe düşürülmüştür. Bu özellik vektörüne de *Vgbtga* adı verilmiştir.

3 Sınıflandırıcı ve Özellik Seçici Algoritmalar

Bu çalışmada, yazarı bilinmeyen bir dokümanın yazarının belirlenmesinde kullanılmak üzere makine öğrenmesi yöntemlerinden dört tanesi seçilmiştir. Bunlar Naïve Bayes, Destek Vektör Makinesi, C 4.5 ve Random Forest'dır. Adı geçen bütün sınıflandırma algoritmaları WEKA ortamında gerçekleştirilmiştir.

Naïve Bayes (NB)
Klasik Naïve Bayes algoritması genelde kelimelerin ve sınıfların birleşik olasılıkları ile bir dokümanın sınıfının belirlenmesinde kullanılır. Bizim çalışmamızda ise özellikler kelimelerin frekansları değildir ve sürekli dağılımlara sahip olduklarından klasik Naïve Bayes yerine George'un [14] çalışmasında önerilen Naïve Bayes versiyonu kullanılmıştır.

Destek Vektör Makinesi (SVM)
Günümüzde performansı sayesinde oldukça popüler olmuş bir metottur. Sınıfları birbirinden ayıran marjini en büyük, doğrusal bir ayırt edici fonksiyon bulunmasını amaçlar. Doğrusal olarak ayrılamayan örnekler için örnekler, doğrusal olarak ayrılabildikleri daha yüksek boyutlu başka bir uzaya taşınır ve sınıflandırma o uzayda yapılır.

C 4.5
C 4.5, tek değişkenli bir karar ağacı algoritmasıdır. Klasik karar ağacı algoritmalarının ana fikri özellik uzayını iki bölgeye ayırmaktır. Bu bölme işlemine her bir bölge içerisinde tek bir sınıfa ait veri kalıncaya kadar devam edilir.

Karar ağaçları, içlerinde verilerin hangi dala yönlendirileceğini belirleyen karar düğümlerinden ve bu dalların uçlarına gelen verinin hangi sınıfta olduğunu söyleyen sınıf etiketlerini içeren yapraklardan oluşan hiyerarşik bir yapıdır. Bir veri, karar ağacıyla sınıflandırılmak istediğinde en tepedeki kök karar düğümünden başlanır ve bir yaprağa gelinceye kadar karar düğümlerindeki yönlendirmelere göre dallarda ilerler. Yaprağa gelindiğinde ise verinin sınıfı yaprağın temsil ettiği sınıf olarak belirlenir. Karar düğümlerinde, eğer ağaç tek değişkenli ise tek bir özelliğin adı ve bir eşik değeri yer alır. O düğüme gelen verinin hangi dala gideceğine verinin o düğümdeki özelliğinin eşik değerinden büyük ya da küçük olmasına göre karar verilir.

Random Forest (RF)
Breiman tek bir karar ağacı üretmek yerine her biri farklı eğitim kümeleriyle eğitilen çok sayıda, çok değişkenli ağacın kararlarının birleştirilmesini önermiştir. Farklı eğitim kümeleri önyükleme (bootstrap) ve rasgele özellik seçimi ile orijinal eğitim setinden oluşturulur. Çok değişkenli karar ağaçları CART [15] algoritmasıyla elde edilir. Önce her karar ağacı kendi kararını verir. Karar ormanı içerisinde maksimum oyu olan sınıf son karar olarak kabul edilir ve gelen test verisi o sınıfa dahil edilir.

Korelasyon Tabanlı Özellik Seçici (Correlation-based Feature Selection-CFS)
CFS, Weka içerisinde yer alan özellik seçici metotlardan biridir. Diğer özelliklerle düşük korelasyonlu, sınıf değişkeni ile yüksek korelasyonlu olan özellikleri seçer.

4 Deneysel Sonuçlar

Deneysel sonuçlarımızı alırken farklı alanlarda yazan 18 yazarın 35 farklı yazısı kullanılmıştır. Her bir yazara ait 10 adet özellik vektörü oluşturulmuştur. Özellik vektörlerinin yazar belirmedeki başarılarını test ederken 10'lu çapraz geçerlilik (10-

fold cross validation) kullanılmıştır. Weka paketi içerisinde yer alan sınıflandırma yöntemlerini kullanırken (Naïve Bayes, SVM, C 4.5 ve RF) öntanımlı (default) parametreler tercih edilmiştir. Bu çalışma Türkçe için yapılan en kapsamlı yazar tanıma çalışmalarından olup, 18 sınıf için oldukça iyi bir sonuç vermiştir.

Çizelge-1'de *Vg, Vbg, Vtg, Vbtg* ve *Vgbtg* vektörlerini yazar belirlemede kullandığımızda alınan başarı sonuçları verilmiştir.

Çizelge 1. Özellik azaltılmadan kullanılan 5 vektörün sınıflandırma başarısı

	(%)Vg	**(%)Vbg**	**(%)Vtg**	**(%)Vbtg**	**(%)Vgbtg**	**(%)ort**
NB	66.5	69.4	70.2	78.1	78.1	71.7
SVM	80	88.1	91.6	92.2	***92.5***	88.9
C 4.5	51.3	56.8	46	61.1	63.5	55.7
RF	48	51.6	42.5	46	45.7	46.8
(%)ort	61.5	66.5	62.6	68.4	70	***65.8***

Kullanılan bu 5 özellik vektöründen sınıflandırmada en büyük başarıyı, hem istatistiksel ve dilbilgisi hem de n-gram modellerinin birleşiminden elde edilen *Vgbtg* vektörü elde etmiştir. Bu vektörün başarısı 4 sınıflandırma yöntemi için ortalama %70'dir. Her bir özellik vektörünün ortalama başarısı *Vg*, *Vbg*, *Vtg*, *Vbtg* ve *Vgbtg* için sırasıyla %61.5, %66.5, %62.6, %68.4 ve %70'dir. Sınıflandırma metotlarının en başarılısı ise büyük bir farkla SVM (%88.9)'dir. En yüksek başarı %92.5 olup *Vgbtg* özellik vektörü kullanılarak SVM ile sınıflandırma yapıldığında elde edilmiştir ve Çizelge 2'de buna ait hata matrisi (confusion matrix) verilmiştir.

Çizelge 2. Vgbtg özellik vektörünün SVM için Hata Matrisi

		Gerçek																	
	Yz	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18
Tahmin edilen	01	30			1				1							1		1	1
	02		35																
	03		1	32			1		1										
	04				32		1												2
	05					32						1						1	1
	06						33				1							1	
	07							33		1					1				
	08			1			2		29		1							1	1
	09									34									1
	10										31						1	3	
	11		1									33						1	
	12				2								33						
	13													34				1	
	14													1	34				
	15		2													32	1		
	16														1		32	2	
	17										3			1	1			30	
	18														1				34

Çizelge-3'te *Vga, Vbga, Vtga, Vbtga* ve *Vgbtga* vektörlerini yazar belirlemede kullandığımızda alınan başarı sonuçları verilmiştir.

Çizelge 3'ü iki farklı açıdan değerlendirebiliriz. Birincisi, kullanılan bu 5 özellik vektöründen sınıflandırmada en büyük başarıyı, hem istatistiksel ve dilbilgisi hem de n-gram modellerinin bir birleşimini özellikleri azaltılarak elde edilen *Vgbtga* vektörü ile elde etmiştir. Bu vektörün başarısı 4 sınıflandırma yöntemi için ortalama %82,5'dir. İkincisi, hangi özellik vektörlerinin hangi sınıflandırıcılarda daha başarılı sonuç verdiğidir. *Vga*, *Vbga* vektörleri Naïve Bayes'de başarılı sonuçlar vermesine rağmen, *Vtga*, *Vbtga* ve *Vgbta* vektörleri SVM sınıflandırıcısında daha yüksek sonuç vermiştir.

Çizelge 3. Özellik azaltılarak kullanılan 5 vektörün sınıflandırma başarısı

	(%)Vga	**(%)Vbga**	**(%)Vtga**	**(%)Vbtga**	**(%)Vgbtga**	**(%)ort**
NB	75.4	78.4	80.2	85.1	85.6	80,9
SVM	70.3	73.3	83.8	88.1	***88.4***	80,8
C 4.5	58.6	63.7	55.9	67.5	74	63,9
RF	69.5	78.3	69	77.6	82	75,3
(%)ort	68.5	73.4	72.2	79.6	82.5	***75,2***

Her bir özellik vektörünün ortalama başarısı *Vga*, *Vbga*, *Vtga*, *Vbtga* ve *Vgbtga* için sırasıyla %68.5, %73.4, %72.2, %79.6 ve %82.5'dir. En yüksek başarı %88.4 olup *Vgbtga* özellik vektörü kullanılarak SVM ile sınıflandırma yapıldığında elde edilmiştir.

5 Sonuç

Bu çalışmada yazarı bilinmeyen bir dokümanın önceden belirlenmiş 18 farklı yazar içerisinden hangisine ait olabileceği sorusunun cevabı aranmıştır. Naïve Bayes, SVM, C 4.5 ve Random Forest sınıflandırıcıları seçilerek 10 ayrı özellik vektörü, 10'lu çapraz geçerlilik ile çalıştırılmış ve oldukça başarılı sonuçlar elde edilmiştir. Yazarlık özellikleri çıkarılırken Türkçe dili için çalışma yapılmıştır. Türkçe'nin istatistiksel, dilbilgisel özelliklerinin yanı sıra Türkçe dilinin n-gram modelleri de çıkarılarak beş farklı özellik vektörü elde edilmiş ve her bir özellik vektörüne özellik seçimi uygulanarak 5 tane daha özellik vektörü oluşturulmuştur. Bu 10 özellik vektörünün sınıflandırma başarısı araştırılmıştır.

En başarılı özellik vektörü; istatistiksel veriler, kelime sözlüğünün zenginliği, sık geçen Türkçe kelimeler, dilbilgisi özellikleri ve n-gram'lardan oluşan *Vgbtg* olmuştur. *Vgbtg*'nin sınıflandırmadaki ortalama başarısı %70 iken, özellik sayısı azaltıldığında (*Vgbtga*) ortalama başarı %82.5'a yükselmiştir.

Özellik sayısı azaltılmamış vektörler (*Vg, Vbg, Vtg, Vbtg* ve *Vgbtg*) kullanıldığında sınıflandırıcıların ortalama başarıları %65.8 iken, azaltılmış vektörler (*Vga, Vbga, Vtga, Vbtga* ve *Vgbtga*) kullanıldığında %75.2'ye yükselmiştir. Ancak en başarılı sonuç özellik sayısı azaltılmamış olan *Vgbtg* vektörünün SVM ile sınıflandırılmasında %92.5 ile elde edilmiştir.

Kullanılan sınıflandırıcıların en başarılısı, orijinal vektörlerle çalışıldığında açık farkla SVM, özellik azaltılmış vektörlerle çalışıldığında ise Naïve Bayes ve SVM'dir.

Sonuç olarak Türkçe dokümanların yazarlarını belirlemede tek başlarına kullanıldıklarında yazarlık özelliklerindense n-gram'ların kullanılmasının daha başarılı olduğu gözlemlenmiştir. Bununla birlikte n-gram ve yazarlık özellikleri birlikte kullanıldıklarında ise her ikisinden de daha başarılı sonuçlar elde edilmiştir. Ayrıca mevcut derlem içerisinde yer alan dokümanların sayısı arttıkça doğru sınıflandırma başarısı yükselecektir, buna karşılık sınıf sayımızı arttırdığımızda ise doğru sınıflandırma başarısı düşecektir.

Çalışmamızın yazarlık özelliklerinin çıkarılmasında, Türkçe için yapılmış en kapsamlı çalışma olduğunu söyleyebiliriz.

Kaynakça

1. Mosteller F., Wallace D.L.: Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Reading, MA:Addison-Wesley (1984)
2. Burrows J.F.: Not unless you ask nicely: the interpretative nexus between analysis and information. Literary Linguist Comput 7:91-109 (1992)
3. Stamatatos E., Fakotakis N., Kokkinakis G.: Automatic Text Categorization in Terms of Genre and Author, Computational Linguistics, pp.471-495 (2000)
4. Morton A.Q.: The Authorship of Greek Prose Journal of the Royal Statistical Society, Series A, 128:169-233 (1965)
5. Brainerd B.: Weighting Evidence in Language and Literature: A Statistical Approach. University of Toronto Press (1974)
6. Holmes D.I.: Authorship Attribution, Comput Humanities, 28:87-106 (1994)
7. Tweedie F., Baayen H.: How Variable may a Constant be Measures of Lexical Richness in Perspective, Computers and the Humanities, 32(5):323-352 (1998)
8. Stamatatos E., Fakotakis N., Kokkinakis G.: Computer-Based Authorship Attribution Without Lexical Measures, Computers and the Humanities, 35: pp.193-214 (2001)
9. Fürnkranz J.:A Study using n-gram Features for Text Categorization Austrian Research Institute for Artifical Intelligence (1998)
10. Tan C. M., Wang Y. F., Lee C. D.: The Use of Bi-grams to Enhance *Jo*urnal Information Processing and Management, Vol:30 No:4 pp.529-546 (2002)
11. Çatal Ç., Erbakırcı K., Erenler Y. : Computer-based Authorship Attribution for Turkish Documents, Turkish Symposium on Artificial Intelligence and Neural Networks (2003)
12. Diri B., Amasyalı M.F.: Automatic Author Detection for Turkish Texts, Artificial Neural Networks and Neural Information Processing, 138-141 (2003)
13. Amasyalı M.F., Diri B.: 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. George H.: Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 338-345. Morgan Kaufmann, San Mateo (1995)
15. Breiman L.: "Random forests–random features," Technical Report 567, Department of Statistics, University of California, Berkeley (1999)