Türkçe Dokümanlarda Yapay SinirAğları ile Yazar Tanıma

Vecdi Emre Levent, Banu Diri
Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul
emre@levent.tc, banu@ce.yildiz.edu.tr

Özet: Bu çalışmada Yapay Sinir Ağları ile Türkçe gazete köşe yazarlarının yazarlık özellikleri çıkarılarak yazar tanıma gerçeklenmiştir. Yazarlık özelliği olarak 16 özellik kullanılmıştır. Hem aynı kategoride hem de farklı kategorilerde yazan yazarlardan ve ayrıca cinsiyete göre üç farklı veri seti kullanılarak yazar tanıma üzerine çalışılmıştır. Yine bu çalışma kapsamında kullanıcı dostu bir Yazar Tanıma uygulaması geliştirilmiştir. Bu uygulama Yapay Sinir Ağlarını kullanarak yazar tanıma yaparken, farklı sayıda sınıflar için oluşturulan özellik setlerini arff dosya formatında da saklamaktadır.

Anahtar Sözcükler: Yapay Sinir Ağları, Yazar Tanıma, Metin Sınıflama, Yazarlık Özellikleri

Author Recognition on Turkish Documents with Artificial Neural Network

Abstract: In this study, realized an author recognition which is Turkish newspaper columnists revealed authoring features with Artificial Neural Networks. 16 features are used as authoring features. Performed author recognition by 3 different datasets which are authors writing on both same category and different category and also by gender. And also an author recognition application developed which is has user friendly interface. This application performs author recognition with Artificial Neural Network, and also saves as arff file format for different number of classes generated feature sets.

Keywords: Artificial Neural Network, Author Recognition, Text Classification, Authorship Feature

1. Giriş

Doküman sınıflandırma ve doğrulama işlemleri doğal dil çalışmaları altında yer alan bir konudur. Eğer amacımız dokümanları türlerine ayırmak ise, kullanacağımız özellikler ile bir dokümanın yazarını bulmak için kullanılan özellikler birbirlerinden farklı olabilirler. Yazar tanıma çalışmaları uzun yıllardır üzerinde çalışılan bir konu olup, Yapay Sinir Ağları, istatistiksel yöntemler ve makine öğrenmesi gibi teknikler yazar tanımada kullanılmıştır. Yazar tanımada başarıyı belirleyen öncelikli olarak seçilmiş olan özelliklerdir. Bu özellikler yazarın stilini belirlemeye yönelik olmalıdır [1]. Yazar tanıma üzerine Türkçe dışında çok fazla çalışma yapılmış olmasına rağmen Türkçe üzerine yapılan çalışma sayısı sınırlıdır.

Bu alanda yapılan öncül çalışmalar 70'li yıllarda, dokümanların otomatik dizinlenmesi olarak karşımıza çıkmaktadır. Çalışmalarda belirli bir konu için özel sözlükler oluşturulmuş, bu sözlüklerin içerisindeki kelimeler sınıf etiketi gibi düşünülerek dokümanların sınıflandırma işlemleri gerçekleştirilmiştir. Mosteller ve Wallace çalışmalarında yazarlık özelliklerini çıkarmış ve tanıma için Bayesian analizi kullanmışlardır [2]. Burrows, dokümanda geçen kelimelerin frekansına [3], Brinegar kelime ve harf sayılarına [4], Morton cümlelerin uzunluklarına [5], Brainerd hece sayılarının ortalama değerine [6], Holmes dokümanda geçen kelime sayısına ve dokümanın uzunluğuna [7], Twedie ve Baayen dokümanda geçen bir adet kelime sayısının dokümandaki toplam kelime sayısına oranını yazar özellikleri olarak kullan-

mışlardır [8]. Stamatatos ve arkadaşları, yazar için bir dizi stil belirleyicisi çıkarmış ve bunları sözlüksel (lexical), karakter, sözdizimsel ve anlamsal olarak gruplandırarak kullanmışlardır [9]. Fürnkranz [10] karakter n-gram'larını (2 ve 3 uzunluğunda), Tan ve arkadaşları da [11] 2-gram'ları (bi-gram) kullanmışlardır. N-gram yöntemi kullanılarak geliştirilmiş olan NECL adındaki sistem Çatal ve arkadaşları tarafından geliştirilmiştir [12]. 22 adet stil belirleyicisi oluşturmuş olan Diri ve Amasyalı ise doküman türü ve yazarının belirlenmesi üzerine sınıflandırma sistemi geliştirmişlerdir [13]. Yine Diri ve Amasyalı 2 ve 3-gram'ları kullanarak yazar tanıma üzerine çalışmışlardır [14, 15]. Amasyalı ve Yıldırım ise Türkçe haber metinlerinin sınıflandırılması üzerine çalışmıştır [16].

Bu çalışmayı referans [14]'te verilen çalışmadan farklı ve güçlü kılan özellik geliştirilen uygulamadaki esnekliktir. Bu sayede kullanıcı arayüzden Yapay Sinir Ağına ait parametreleri değiştirebilmektedir. Toplamda 16 adet olan yazarlık özelliklerinin çıkarılmasında Zemberek kütüphanesinden yararlanılmıştır [17]. Geliştirilen sistem oldukça esnek, parametrik ve kullanıcı dostu olarak tasarlanmıştır. Aynı zamanda farklı veri setleri ve farklı yazar özellikleri ile sistem eğitilirken ilgili bilgiler istenildiği taktirde Weka [18] ile kullanılabilmesi için arff (attribution relation file format) formatında da saklanabilmektedir. Tasarlanan sistem .Net ortamında C# ile geliştirilmiştir.

Makalenin ikinci bölümünde yazarlık özelliklerinden, üçüncü bölümde sınıflandırma için kullanılan Yapay Sinir Ağından, dördüncü bölümünde geliştirilen uygulamanın adımlarından bahsedilmiştir. Beşinci bölümde deneysel sonuçlara yer verilmiştir.

2. Yazarlık Özellikleri

Güncel olarak takip ettiğimiz bir köşe yazarının bir müddet sonra ismi olmadan ve hatta yazısı farklı bir sayfada yayınlandığında bizler yazıyı kimin yazdığını tahmin edebiliriz. Çünkü takip ettiğimiz yazarın yazı üslubuna alışmışızdır. Bizim yapabildiğimiz bu ayırımı bilgisayarlara otomatik olarak yaptırabilir miyiz? Bu sorunun cevabı elbette ki evet olacaktır. Her yazarın kendine özgü bir üslubu vardır ve bu üslup matematiksel bir fonksiyona dönüştürüldüğünde yazar tanıma işlemi gerçekleştirilmiş olur. Yazarların kendilerine özgü olan bu tarzı, yazarlık özelliği (authorship attribution) olarak adlandırılmaktadır. Bu çalışmada her yazar için 16 adet özellik belirlenmiştir. Bu özellikler sırasıyla: cümle ve kelime sayısı, ortalama kelime ve farklı kelime sayısı, nokta, virgül, satır, noktalı virgül, soru işareti, ünlem, isim, fiil, sıfat, zamir, edat ve bağlaç sayısıdır. Kullanıcı sistemin eğitimi sırasında istediği özelliği ekleme veya çıkartma esnekliğine sahiptir. Yapısal sözcükler olan isim, fiil, sıfat, zamir, edat ve bağlaçların tespitinde açık kaynak kodlu Zemberek Kütüphanesinin .Net ortamı için geliştirilmiş olan Nzemberek [17] kütüphanesi kullanılmıştır (Şekil 1).



Şekil 1. Özellik çıkarımı yapısı

Çıkarımı yapılmış olan özelliklerin büyüklükleri birbirinden farklıdır. Örneğin, dokümanda geçen kelime sayısı ile bağlaç sayısının büyüklük değerleri birbirlerine göre oldukça farklı olacaktır. Eğer, sisteme eğitim için özellik değerleri normalize edilmeden verilirse, büyüklüğü az olan özellikler belirleyici olsalar da sisteme etkileri çok az olacaktır. Dolayısıyla normalizasyon işlemi yüksek önem arz etmektedir ve her özellik için bulunan değerler [0,1] aralığına çekilmektedir. Böylelikle tüm özellik değerleri sistem için aynı ayırt edicilik ile değerlendirilecektir.

3. Yapay Sinir Ağları

Beyindeki bilgi işleme yönteminden esinlenerek tasarlanmış olan Yapay Sinir Ağları (YSA),

katmanlar ve bu katmanların içerisinde yer alan nöronlar ile bir ağ oluştururlar. Nöronlara atanan sayısal değerler ve katmanların çıkışlarındaki transfer fonksiyonları ile sistem içerisinde hesaplama yapılır ve çıkış katmanından sonuç alınır. Eğitim esnasında önceden belirlenmiş nöron, katman sayıları ve transfer fonksiyonları ile nöronların değerleri değiştirilerek sonuca olan etkisine bakılır. Hedef her döngüde olması gereken ile hesaplanan değer arasındaki hatayı minimize etmektir.

3.1 Eğitimin Gerçekleştirilmesi

Önceden belirlenmiş, eğitime girecek yazarlar ve yazarlara ait dokümanlar sisteme verilirler. YSA yapısı, hata oranı istenen değerin altına ininceye kadar veya belirli bir döngü sayısı tamamlanıncaya kadar devam ederek hatayı minimize etmeye çalışır. İşlem sonucunda her katmandaki nöronların değeri optimize edilmiş olacaktır. Geliştirilen uygulamada girişler ve olması gereken değerler verilip, hata ve döngü sayısı kontrol edilerek devam eden bir yapı oluşturulmuştur.



Şekil 2. YSA parametreleri

Şekil 2'de örnek bir YSA modeli gösterilmektedir. Özellik sayısına karşılık gelen 16 giriş, 15 nöronlu bir ara katman ve dört farklı yazar için dört çıkıştan oluşan bir ağdır. Girişteki özellik sayısı, ara katman sayısı ve transfer fonksiyonu parametreleri kullanıcı tarafından seçimliktir. Transfer fonksiyonu olarak kullanıcıya sigmoid, lineer, gauss ve oransal sigmoid seçimleri sunulmaktadır.

3.1.1 Transfer Fonksiyonu

Transfer fonksiyonları nöronlardan gelen değerleri toplar. Birleştirip, kullandığı fonksiyon cinsine göre dönüştürerek sayısal değer üretirler. Katmanların çıkışında bulunurlar, her katmanın çıkış değeri burada hesaplanmaktadır.

3.1.2 Öğrenme Oranı

Öğrenme oranı yapay sinir ağlarının eğitiminde kullanılan bir parametredir. Daima sıfıra yakın değerler seçilmesi tavsiye edilmektedir. Seçilen değer pozitif olduğunda eğitimin hedeflenen değere hızlı erişmesine ve sıfır'a yakın küçük değerler seçildiğinde sistem sapma ihtimali azalmaktadır. Sistemin performansı açısından öğrenme oranının önemli bir etkisi vardır. Küçük öğrenme oranı seçilir ise, eğitim sonuçları daha iyi olacaktır ancak öğrenme yavaş ilerleyecektir. Büyük değerler seçilmesi durumunda eğitim daha kısa sürede gerçekleşebilir ancak aranan minimum hata değerine ulaşma ihtimali her zaman yüksek değildir. Çok katmanlı bir yapıya sahip YSA'da yerel minimum hata çözümüne takılabilir ve sistem salınım da kalabilir. Öğrenmenin yavaş olması zaman açısından da bir dezavantaj olduğu için zaman ve minimum hata dengesini kurmak gerekmektedir [19].

3.1.3 Momentum Katsayısı

Momentum katsayısı, ağın yerel minimum değerlerine takılmamasını sağladığı gibi aynı zamanda hatanın düşmesine de yardımcı olur. Dolayısıyla sistem salınıma girmekten korunmuş olmaktadır. Momentum değeri 0 ile 1 arasındadır. Momentum parametresi sisteme verildiğinde, ağırlıklara göre eğim ardışık iki döngü aynı işaretli ise ağırlıklara uygulanacak düzeltme artar, aksi halde azalır. Momentum değerinin sıfır olduğu durumlarda ağırlık değişimleri gradiyente bağlı olmaktadır. Bu değerin bir olması durumunda ise gradiyentten tamamen bağımsız olmaktadır [20].

3.2 Testin Gerçekleştirilmesi

Eğitimin sonunda sistemin parametreleri optimize edilmiş ve ağaç yapısı elde edilmiştir. Sistemin başarısının ölçümünün yapılabilmesi için eğitime verilmemiş dokümanlardan oluşan veri seti sisteme giriş olarak verilmektedir. Çıkış olarak alınacak nöronlar eğitime sokulan

yazar sayısı kadar olacaktır. Buradaki yaklaşım hangi nörondan en yüksek değer geri döner ise, o nörona ait yazar, sisteme verilen dokümanın yazarı olarak tespit edilmiş olacaktır.



Şekil 3. Hata Matrisi

Şekil 3'te örnek olarak sistem altı yazar ile eğitildikten sonra, yazarların eğitime girmemiş on dokümanı ile test edilmiş ve sonuçlar hata matrisi üzerinde gösterilmiştir. Sistemin başarısı tutturma (precision) ve çağrı (recall) nın harmonik ortalaması alınarak hesaplanan f-ölçüm üzerinden verilmektedir.

4. YSA ile Yazar Tanıma Uygulaması

Geliştirilen uygulamanın kurulumuna Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümünün araştırma gruplarından Kemik[1]'in sayfasından ulaşılmaktadır. .Net ortamında geliştirilen uygulama kullanıcıların sistemde geçen tüm parametreleri düzenleyebileceği, mevcut yazarlara yeni dokümanlar ve sisteme yeni yazarlar ekleyebilecek esneklikte tasarlanmıştır.

Mevcut sistemin kullanımında 50 farklı yazar, her yazara ait 50 doküman bulunmaktadır. Dokümanların tamamı Kemik'in sayfasından indirilebileceği gibi kurulum içerisinde de yer almaktadır. Deneysel çalışmalarda sistemi 50 yazarla eğitmek elimizdeki donanım ile oldukça uzun süreceğinden maksimum 16 yazar ile çalışılmıştır. Yazarlar www.tumkoseyazilari.com adresinden seçilmiştir ve burada belirtilen yazar kategorileri kullanılmıştır. Deneysel çalışmalarımız iki farklı bilgisayarda gerçekleştirilmiştir. Kullanılan birinci donanımın özellikleri, AMD A8-5550M APU 2.10 GHz işlemci, 4 GB hafıza, 64 bit işletim sistemi Windows 8.1'dir. Veri Seti-1 (VS1) bu donanım üzerinde çalıştırılmıştır. Veri Seti-II (VS2) ve Veri Seti-III (VS3) ise farklı bir donanım Intel(R) Core(TM)2 Duo CPU P8700 2.53GHz işlemci ve 4GB ram üzerinde çalıştırılmıştır. Uygulamanın ara yüz ekranı Şekil 4'deki gibidir.



Şekil 4. Uygulamanın Başlangıç Arayüzü

5. Deneysel Sonuçlar

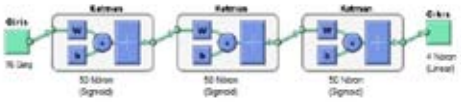
Yapay Sinir Ağları kullanılarak gerçekleştirilen yazar tanıma çalışmasında üç farklı veri seti oluşturulmuştur. Birinci veri seti VS1 olup, farklı kategorilerde yazan 16 farklı yazarın 50 adet gazete makale yazısından oluşmaktadır. Her bir yazarın makalelerinin 40 tanesi eğitim, 10 tanesi de test için kullanılmıştır. Bu yazarların sekiz tanesi güncel, üç tanesi yaşam, iki tanesi spor ve üç tanesi de kültür-sanat kategorisinde yazmaktadır. İkinci veri seti VS2 olup, cinsiyete göre oluşturulmuş veri setidir. İçerisinde 10 kadın, 10 erkek yazardan oluşan ve toplamda 400 makaleden oluşan bir veri setidir. Makalelerin 320 tanesi eğitim için kullanılmış 80 tanesi de test olarak ayrılmıştır. Üçüncü veri seti VS3 olup, aynı kategoride (güncel) yazan 16 yazardan oluşmaktadır. Herbirine ait 40 doküman eğitim, 10 dokümanda test için kullanılmıştır. Deneysel sonuçlarımız her veri seti için dört, sekiz, oniki ve onaltı yazar için alınmıştır.

5.1. Eğitimde Kullanılan YSA Yapısı

Eğitimde kullanılan Yapay Sinir Ağı, çıkarılan 16 adet yazar özelliğinin hepsini kullandığımızdan dolayı 16 girişli, herbiri sigmoid transfer fonksiyonla eğitilen 50 nöronlu üç ara katmandan ve lineer transfer fonksiyonuna sahip sınıf

1 http://www.kemik.yildiz.edu.tr/

sayımız kadar çıkış içeren bir ağ yapısındadır. Şekil 5'te, 16 sınıflı bir veri için oluşturulmuş yapay sinir ağı yapısı görülmektedir.



Şekil 5. YSA Modeli

Yapay Sinir Ağının parametreleri olan öğrenme oranı ve momentum değerleri denemelerde 0,1 olarak alınmıştır. Minimum hata ve iterasyon sayısı kullanıcı tarafından değiştirilebilmektedir. Tüm denemelerimizde minimum hatanın 0,1 olması istenmiştir. Buna karşılık iterasyon sayı olarak 30 bin gibi maksimum bir sayı verilmiştir. Sistemin eğitimi sırasında elde edilen özellik değerleri seçimlik olarak arff (attribution relation file format) olarakta saklanmaktadır. Eğitim sonunda elde edilen ağ, test aşamasında kullanılmak üzere sistem tarafından .xml formatında tutulmaktadır. Sistemin testi sırasında ilgili .xml dosyası ile eğitilen ağ sisteme verilmekte ve test dokümanları ile sistem test edilmektedir. Teste giren dokümanların özellik değerleri de arff dosya formatı olarak saklanmaktadır. Elde edilen arff dosyaları Weka [18] aracı içerisindeki tüm sınıflandırma algoritmaları ile çalıştırılabilinir. Denemelerde 16 özelliğin hepsi ile eğitim gerçekleştirilmiştir. Özellik azaltmaya gidilmemiştir. Çünkü Yapay Sinir Ağlarının giriş parametreleri yazarın üslubunu tanımada hangi özellik daha çok etkili ise o parametrelerin değerlerini yükselterek, az olanların ise düşürerek sistemi eğitmektedir.

5.2. Eğitim Sonuçları

İlk denememiz farklı kategorilerde yazan yazarlardan oluşturulmuş veri seti VS1 ile yapılmıştır. Dört, sekiz, oniki, onaltı yazarın 0,1 minimum hatayla kaç iterasyonda ve ne kadar zamanda eğitildiği ve test verileri ile sistem çalıştırıldığında sistemin başarısının f-ölçüm cinsinden değeri Tablo 1'de verilmektedir. Sınıf sayısı arttıkça başarının düştüğü gözlemlenmektedir. 12 yazarlının başarısının sekiz yazarlıdan daha yüksek olmasının sebebi eklenen dört yazarın yazarlık özelliklerinin daha ayırt edici olmasıdır.

#yazar	F-ölçüm	Hata Oranı	#döngü	Süre (dak)
4	1	0,09	2.792	3.28
8	0,78	0,09	11.119	27.46
12	0,95	0,99	11.619	44.37
16	0,73	0,1	13.768	71.58

Tablo 1. VS1 Deneme Sonuçları

İkinci denememiz cinsiyet üzerine VS2 isimli veri seti ile yapılmıştır. Bu denemede amacımız yazarların stillerinin cinsiyete göre bir farklılık gösterip göstermediğini ortaya koymaktır. Tablo 2'de VS2 ile alınan sonuçlar verilmektedir.

F-ölçüm	Hata Oranı	#döngü	Süre (dak)
0,75	0,16	35000	204.44

Tablo 2. VS2 Deneme Sonuçları

Cinsiyet belirlemede test için kullanılan 10'u erkek ve 10'u kadın yazara ait olan veri setinde, erkeklerin doğru tespit edilme oranı daha yüksek çıkmıştır. Buradan erkeklerin yazma karakteristiklerinin birbirlerine daha yakın olduğu, kadınların ise daha farklı yazma karakteristiklerine sahip olduğu sonucuna varılmıştır.

Üçüncü denememiz VS3 isimli veri seti ile yapılmıştır. Bu veri seti içerisinde yer alan yazarların hepsi "Güncel" olarak isimlendirilen alanda yazan yazarlardan seçilmiştir. Bu veri seti ile yapılan çalışmanın amacı aynı alanda yazan yazarların yazı içerikleri birbirlerine yakın olmasına rağmen yazarlık stillerinin ne ölçüde ayırt edici olduğunu göstermektir. Tablo 3'te VS3 ile alınan sonuçlar verilmektedir.

VS3'in başarısı VS1'e göre daha iyidir. Aynı konuda yazan yazarların stilleri birbirlerine benzerlik göstermediğinden sınıflar daha düzgün ayrılabilmişlerdir.

#yazar	F-ölçüm	Hata Oranı	#döngü	Süre (dak)
4	0,98	0,1	5334	10.41
8	0,98	0,1	12876	51.0
12	0,98	0,1	14802	86.29
16	0,83	4.18	50000	262.2

Tablo 3. VS3 Deneme Sonuçları

YSA ile aldığımız sonuçların makine öğrenmesi yöntemlerinden biriyle karşılaştırmak istedik. Bu amaç ile geliştirilen uygulama ile üretilmiş olan arff formatındaki dosyaları kullanarak, Weka uygulaması altında yer alan SVM (Support Vector Machine) yöntemini seçip, çalıştırdık.

İlk veri setimiz olan farklı kategorilerde yazan yazarlara ait veri seti VS1, SVM ile çalıştırıldığında alınan sonuçlar ve YSA ile karşılaştırılması Tablo 4'te verilmiştir.

#yazar	Yazar Tanıyıcı (YSA F-Ölçüm)	Weka- SVM (F-Ölçüm)
4	1	0,95
8	0,78	0,75
12	0,95	0,84
16	0,73	0,75

Tablo 4. VS1 YSA-SVM Karşılaştırılması

Cinsiyete bağlı olan, ikinci veri setimiz VS2 ile SVM çalıştırıldığında karşılaştırmalı sonuçları Tablo 5'te görebilirsiniz.

Yazar Tanıyıcı (YSA F-Ölçüm)	Weka - SVM (F-Ölçüm)
0,75	0,75

Tablo 5. VS2 YSA-SVM Karşılaştırılması

Cinsiyete bağlı olan veri setinde her iki yöntemde aynı başarıyı vermiştir. Üçüncü veri setimiz olan aynı kategoride yazan yazarlardan oluşan VS3 ile SVM çalıştırıldığında her iki yöntemin karşılaştırmalı sonuçları Tablo 6'da verilmektedir.

#yazar	Yazar Tanıyıcı (F-Ölçüm)	Weka-SVM (F-Ölçüm)
4	0,98	1
8	0,98	0,92
12	0,98	0,84
16	0,83	0,88

Tablo 6. VS3 YSA-SVM Karşılaştırılması

VS1 veri setinde olduğu gibi VS3 veri setinde de 16 sınıfta SVM daha iyi sonuç vermektedir.

Deneysel çalışmalarda YSA ile sistemi eğitirken eğitim uzun sürmekte ancak sistem bir kez eğitildikten sonra test işlemi çok hızlı yapılabilmektedir. 2'den 16'ya kadar tüm sınıflarda hep aynı ağ modelini kullandığımız için az sınıf sayılarında sistem gereksiz yere çalışmıştır.

6. Sonuçlar ve Öneriler

Yazar tanıma konusunda farklı dillerde çalışmalar mevcuttur. Türkçe üzerine de kısıtlı sayıda yapılmış çalışma mevcuttur. Bu çalışmada kullanılan yazarlık özellikleri diğer çalışmalarda kullanılmış olan özelliklerle de benzerlik gösterir. Ancak, bu çalışma için geliştirilmiş uygulamayı bölüm 4' de belirtilen adresten kolayca indirebilir ve masaüstü uygulaması olarak kullanabilirsiniz. Yeni yazar ekleme veya mevcut yazarlara doküman ekleme, belirli özellikleri seçme, kullanılan yapay sinir ağının parametrelerini değiştirebilme imkanı kullanıcıya verilmiştir. Yine yazar tanıma çalışmalarının çoğunda kullanılan bir veri madenciliği aracı olan Weka da kullanılmak üzere arff dosya formatında veri setleri de oluşturulmaktadır.

Çalışmanın ilerletilmesi bakımından paralel işleme özelliği olan fpga, gpu gibi cihazlar üzerinde ağın eğitimi yapıldığı taktirde işlem çok daha verimli olacaktır.

7. Kaynaklar

[1] Türkoğlu, F., Diri, B., Amasyalı, F., "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", Turkish Symposium on Artificial Intelligence and Neural Networks, Turkey (2006)

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

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

[4] Brinegar, C.S., "Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship", Journal of the American Statistical Association, 58:85-96 (1963)

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

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

[7] Holmes, D.I., "Authorship Attribution", Comput Humanities, 28:87-106 (1994)

[8] 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)

[9] Stamatatos, E., Fakotakis, N., Kokkinakis, G., "Computer-Based Authorship Attribution Without Lexical Measures", Computers and the Humanities, 35: pp.193-214 (2001)

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

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

[12] Çatal, Ç., Erbakırcı, K., Erenler, Y., "Computer-based Authorship Attribution for Turkish Documents", Turkish Symposium on Artificial Intelligence and Neural Networks (2003)

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

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

[15] Amasyalı, M.F., Diri, B., "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender", NLDB, 221-226, (2006)

[16] Amasyalı, M.F., Yıldırım, T., "Otomatik Haber Metinleri Sınıflandırma", SIU, 2004.

[17] https://code.google.com/p/nzemberek/ Erişim tarihi: Kasım 2013

[18] http://www.cs.waikato.ac.nz/ml/weka/ Erişim tarihi: Kasım 2013

[19] Elmas, Ç., Yapay Sinir Ağları, Seçkin Yayınevi, Ankara (2003)

[20] Sağıroğlu, Ş., Beşdok, E., Erler, M. Mühendislikte Yapay Zeka Uygulamaları-I Yapay Sinir Ağları, Ufuk Yayınevi, Kayseri (2003)