DÜMF Mühendislik Dergisi
web: http://dergipark.gov.tr/dumf

Türkçe Köşe Yazılarında Yapay Sinir Ağlarıyla Yazar ve Gazete Tahmin Etme

Emrah Aydemir*
Kırşehir Ahi Evran Üniversitesi, Bilgisayar Mühendisliği Bölümü, Kırşehir
emrah.aydemir@ahievran.edu.tr ORCID: 0000-0002-8380-7891, Tel: (0386) 280 38 00 (6056)

Geliş: 21.05.2018, Kabul Tarihi: 31.05.2018

Öz

Doğal dil işleme alanı doküman sınıflandırma ve doğrulama işlemleri ile ilgilenmektedir. Bir metnin yazarı tespit edilmek istenirse kuşkusuz en önemli unsur kullanılacak özelliklerdir ve bu özellikler doğrudan başarıya etki edecektir. Bu çalışmada dört farklı Türkçe gazetenin her birinden 10 adet yazar rastgele olarak seçilmiş ve her bir yazarın da toplam 10 adet köşe yazısı rastgele tespit edilmiştir. Yazarı tanımaya yönelik olarak belirlenen 30 adet özellik yazar tanıma için belirlenmiş ve geri yayılımlı yapay sinir ağlarına girdi olarak verilmiştir. Çıktı olarak ise yazar adı modelinin kurgulandığı bu çalışmada eğitim ve test verileri altı farklı şekilde k-katlı çapraz doğrulama yöntemi ile ayrıştırılmıştır. İç katmandaki sinir sayıları da farklı katman ve değerlerde değiştirilerek denemeler yapılmış ve en iyi modele ulaşmak hedeflenmiştir. Çalışma sonucunda her bir gazete için farklı doğruluk oranları elde edilmiştir. En yüksek başarı oranı %86.9 iken, en düşük başarı oranı %75.0 elde edilmiştir. Başarı oranlarının birbirlerinden farklı çıkmasında ise her gazetedeki yazarın yazarlık özelliklerinin ayırt ediciliği etkili olduğu düşünülmektedir.

Anahtar Kelimeler: *Makine Öğrenmesi, Yazar Tahmini, Metin Sınıflandırma*

* Yazışmaların yapılacağı yazar

DOI: 10.24012/dumf.425754

Giriş

Bilişim alanında kullanılan sistemlerin sayısı arttıkça depolanan veri sayısı da bu doğrultuda artmaktadır. Bilhassa metinlerin benzerliklerine göre sınıflandırılması ve yazarının belirlenmesi önemli sorunlar arasında görülmektedir. Metinlerin sınıflandırılmasının temel amacı o metinin özellikleri dikkate alınarak daha önceden belirlenen kategoriler arasından hangisine ait olduğunu belirlemektir. Bu şekilde yapılan metin sınıflandırma işlemlerinde bilgi elde etme, indeksleme, filtreleme, hiyerarşik düzenleme imkânı sağlar (Amasyalı, Diri ve Türkoğlu, 2006). Bu tür duruma en iyi örnek olarak e-posta içeriklerinin analiz edilerek gereksiz (spam) olup olmadığına karar vermek veya arama motorları ile en iyi ilgili sıralamayı elde etmektir.

Doğal dil alanında yapılan çalışmalar ağırlıklı olarak metin sınıflandırma konuları ile ilgilenmektedir. Burada belirgin ayırt edici durum seçilen özellikler olmaktadır. İstenen metin türlerine göre sınıflandırılmak istenmesi durumu ile yazarını belirlemek için yapılan sınıflandırmada seçilen özellikler birbirinden farklı olacaktır. Elbette bu tür sınıflandırma çalışmalarında başarı oranını doğrudan etkileyen en büyük etmen seçilen özellikler olacaktır. Belirlenen tüm özellikler doğrudan ilgili sınıflandırma ile ilgili olmak zorundadır. Farklı dillerde sınıflandırma ve yazar belirleme çalışmaları var iken Türkçe diline ait çalışmalar maalesef daha sınırlı sayıdadır. 70'li yıllarda metinlerin etiketlenmesi üzerine öncül çalışmalar görülmüştür (Levent ve Diri, 2014). Fakat Türkçe diline ait çalışmalara ise 1999 yılında başlanmıştır.

Metinlerin Belirleyici Özellikleri

Bir metnin yazarına ait birden fazla farklı metinler incelenir ve analiz edilirse bu metinlerin birbiri ile bazı benzerlikler taşıdığı görülecektir. Bu benzerlikler ayrıca metnin yazarını da tahmin edilmesine imkân tanıyacaktır. Bilhassa gazete metinlerindeki köşe yazılarının incelenmesi ve analiz edilmesi sonrası yazarı daha rahat belirlenebilir. Her yazarın yazım üslubu istatistiksel verilere yansıtılması ile bilgisayar tarafında kullanılacak belirli algoritmalar ile yazar tahmini yapılabilir. Metin sınıflandırma çalışmalarında k-en yakın komşu, naive bayes, destek vektör makineleri, J48, rastgele orman yöntemleri vb. farklı yöntemler kullanılmıştır (Fung ve Mangasarian, 2003; Diri ve Amasyalı, 2003; Aşlıyan ve Günel, 2011; Soucy ve Mineau, 2001). Bir metnin yazarını belirlemek için çıkartılan özelliklere yönelik benzer çalışmalar incelendiğinde kelime frekanslarına, harf sayılarına, cümle uzunluklarına, ortalama hece sayılarına, metnin toplam uzunluğuna, kelimelerin tüm kelimelere oranına bakılmıştır (Burrows, 1992; Brinegar, 1963; Morton, 1965; Holmes, 1994; Tweedie ve Baayen, 1998).

Yazar tahmin etmede başarıyı etkileyen unsurlar arasında en önemlisi seçilen özelliklerdir. Fakat bu özelliklerin yanı sıra dil, seçilen metinler ve yazarlar da doğrudan veya dolaylı olarak başarıyı etkilemektedir. Literatürde yapılan çalışmaların başarı oranı incelendiğinde yazar sınıflandırmada %74 oranında doğru sınıflandırma yapan çalışmalar var iken %97 oranında doğru sınıflandırma yapan çalışmalar da görülmektedir. Fakat bu oranlar arasında yer alan %80, %81, %83, %85, %90 oranında başarının elde edildiği çalışmalar da bulunmaktadır (Cavnar ve Trenkle, 1994; Stamatatos, Fakotakis ve Kokkinakis, 2000; Peng ve Schuurmans, 2003; Amasyalı ve Diri, 2006; Peng, Wang ve Schuurmans, 2003).

Materyal ve Yöntem

Veri Toplama

Bu çalışmada hem basılı hem de internet üzerinden yayın yapan dört farklı gazetenin her birinden 10 adet köşe yazarı tespit edilmiştir. Her yazarın ise toplam 10 tane köşe yazısı kayıt altına alınmıştır. Toplamda 400 adet köşe yazısı elde edilmiştir. Bu yazılar Yazar Adı, Gazete, Köşe Yazısı sütunlarından oluşan bir veritabanı tablosunda kaydedilmiştir.

Önişlem

Veri kümesindeki tüm kayıtlar yazar tarafından geliştirilen .net tabanlı bir programa Zemberek kütüphanesini kullanılarak sunulmuştur. Program aracılığıyla metinlerin 30 adet özelliği belirlenmiş ve geri yayılımlı yapay sinir ağlarına girdi olarak sunulmuştur. Çıktı olarak ise yazar adı ile gazete adı verilmiştir. Aşağıda Weka programında girdi olarak kullanılan nitelikler verilmiştir.

- Cümle Sayısı
- Harf Sayısı
- Paragraflardaki Ortalama Cümle Sayısı
- Ortalama Kelime Uzunluğu
- Cümledeki Ortalama Kelime Sayısı
- Kelime Sayısı
- Farklı Kelime Sayısı
- Nokta Sayısı
- Virgül Sayısı
- Karakter Sayısı
- Paragraf Sayısı
- Noktalı Virgül Sayısı
- Soru İşareti Sayısı
- Ünlem Sayısı
- İsim Kelime Türü Sayısı
- Özel İsim Kelime Türü Sayısı
- Sıfat Kelime Türü Sayısı
- Fiil Kelime Türü Sayısı
- Zamir Kelime Türü Sayısı
- Bağlaç Kelime Türü Sayısı
- Edat Kelime Türü Sayısı
- Sayı Kelime Türü Sayısı
- Zaman Kelime Türü Sayısı
- Soru Kelime Türü Sayısı
- Bilinmeyen Kelime Türü Sayısı
- Kısaltma Sayısı
- Çift Tırnak Sayısı
- Tek Tırnak Sayısı
- Tire Sayısı
- Parantez Sayısı

Sınıflandırma ve Başarı Ölçütü

Girdi değerlerinin sayısal veya metin türlerinden oluşması ve çıktı değerinin ise yalnızca kategorilendirilmiş metinlerden oluşması durumunda metin sınıflandırması yapılır. Weka programı aracılığıyla metin sınıflandırma yöntemlerinden herhangi biri kullanılabilir. Programın varsayılan kurulumu ile BayesNet, NaiveBayes, NaiveBayesMultiNominalText, NaiveBayesUpdateable, LibSVM, Logistic, MultiLayerPerceptron, SGD, SGDText, SimpleLogistic, SMO, VotedPerceptron, IBk, kStar, LWL, DecisionTable, JRip, J48, RandomForest vb. birçok yöntem kurulu gelmektedir ve kullanılabilmektedir. İstenirse paket yöneticisi aracılığıyla başka yöntemlerin de kurulumu yapılabilir. Bu algoritmalar aşağıdaki gibi gruplandırılabilir.

- Bayes Sınıflandırıcılar
- Ağaç Algoritmaları
- Kural Tabanlı Sınıflandırıcılar
- Fonksiyonlar
- Tembel Algoritmalar
- Meta Öğrenme Algoritmaları
- Çeşitli Sınıflandırıcılar

Buradaki çalışmada her bir yöntemin detaylarına girmeden çalışma içerisinde kullanılan yöntemler hakkında kısaca bilgi verilecektir. Bu çalışmada her bir gruptan bir yöntem belirlenmiş ve yöntemin sonuçları başarı oranı açısından karşılaştırılacaktır.

NaiveBayesUpdateable

NaiveBayes algoritması olasılıksal temel bayes sınıflandırıcıyı temel alarak olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile sunulan verilerin sınıfını tespit etmeye çalışır. NaiveBayesUpdateable algoritması ise NaiveBayes'e göre her seferinde bir örneği işleyen artımlı bir sürümdür. Bir çekirdek tahmincisi kullanabilir, ancak ayrıklaştırma yapamaz.

MultiLayerPerceptron

İnsan beyninin çalışma ilkesinden esinlenerek geliştirilmiş, her biri belirli ağırlıklara sahip bağlantılar aracılığıyla birbirine bağlanan ve yine her biri kendi belleğine sahip işlem elemanlarından oluşan paralel ve dağıtılmış bilgi işleme yapılarına yapay sinir ağları denir

(Malikoğlu, 2002). Bu algoritma Weka programı içinde yapay sinir ağları algoritmasını uygular ve çeşitli parametreler varsayılan olarak gelmekle birlikte kullanıcı tarafından değiştirilmesine izin verir.

IBk

Tembel öğrenme algoritmaları eğitim örneklerini saklarlar ve sınıflandırma zamanına kadar gerçek bir çalışma yapmazlar. En basit tembel öğrenme algoritması IBk olarak adlandırılan k-en yakın komşu sınıflandırıcısıdır. En yakın komşu bulma görevini hızlandırmak için çeşitli farklı arama algoritmaları kullanabilir.

AdaBoostM1

Meta öğrenmeler basit sınıflandırma algoritmalarını alır ve onları daha güçlü öğrenme algoritmalarına dönüştürür. Önceki modelin yanlış sınıflandırdığı örnekleri vurgulamak için her yeni modeli eğiterek işlem yapar. Bu algoritma ile örneğin iç içe yapay sinir ağları (multilayerperceptron) algoritması kullanılabilir.

InputMappedClassifier

Bir temel sınıflandırıcıyı (ya da bir dosyaya serileştirilmiş olan modeli) sarar ve gelen test verilerinde bulunan özellikler ile model eğitildiğinde görülen özellikler arasında bir eşleme oluşturur. Test verilerinde bulunan, ancak eğitim verilerinde bulunmayan nitelikler için değerler basitçe göz ardı edilir. Eğitim verilerinde olan, ancak test verilerinde bulunmayan nitelikler, eksik değerleri alır. Benzer şekilde, eksik değerler eğitim verilerinde olmayan yeni nominal değerler için kullanılır.

DecisionTable

Bir karar tablosu sınıflandırıcısı oluşturur. En iyi ilk arama özelliğini kullanarak özellik alt kümelerini değerlendirir ve değerlendirme için çapraz doğrulamayı kullanabilir. Aynı özellik kümesine dayanarak, tablonun global çoğunluğu yerine, bir karar tablosu girdisiyle kapsanmayan her bir örnek için sınıfı belirlemede en yakın komşu yöntemini kullanır.

J48

Kısmi karar ağaçlarından kurallar alır. C4.5'in sezgisel özelliklerini kullanarak ağacı J48 ile aynı kullanıcı tanımlı parametrelerle oluşturur.

Başarı Oranı Analizi

Kategorik sınıf değerleri tahmin edilirken amaç kaç tane türden kaç tanesinin doğru sınıfa yerleştirildiğini tahmin etmektir. Buradaki sınıflardan hangilerinin hangi sınıfa yerleştirildiğini görmek için hata matrisi (Confusion Matrix) tablosu kullanılır. Hata matrisi tablosu incelenerek test verilerinden kaç tanesinin doğru ve yanlış yerleştirildiği sonucu elde edilebilir.

Kappa istatistiği satır ve sütun sayısı eşit olan tablolarda iki değişken arasındaki uyumu ölçmek için kullanılır. Kolay hesaplanıp pratik olarak yorumlanabilen şans ile beklenen arasındaki uyumu düzeltmeyi temel alır. Kappa istatistiği, -1 ile +1 arasında değerler alır ve 0'dan küçük olması durumunda uyum olmadığı 1'e yaklaştıkça ise tam bir uyumun olduğunu gösterir. Kappa istatistiği hesaplanırken iki farklı olasılık hesaplanır. Bunlar Pr(a) ve Pr(e)'dir. Pr(a) iki değerlendirici için gözlemlenen uyumların toplam orantısı iken, Pr(e) bu uyumun şansa bağlı ortaya çıkma olasılığıdır. Bu iki olasılık üzerinden Cohen'in kappa istatistiği için kullanılacak formül aşağıdaki gibidir (Sim ve Wright, 2005).

$$K = \frac{\Pr(\mathrm{a}) - \Pr(e)}{1 - \Pr(e)}$$

Tahmin edilen sınıf kategorik olması sebebiyle her bir kategorinin ne kadar doğru sınıflandırıldığının detaylı analizi görülebilmektedir. Bunun için öncelikle ikili bir sınıflandırma tahmininde hata matrisindeki değerleri doğru anlamak gerekir. Öncelikle aşağıdaki tanımlar verilmiştir.

- **Doğru Pozitif:** Gerçek değeri pozitif olup pozitif olarak tahmin edilenler. Türkçe *dp* olarak kısaltılacaktır.
- **Yanlış Negatif:** Gerçek değeri pozitif olup negatif olarak tahmin edilenler. Türkçe *yn* olarak kısaltılacaktır.
- **Yanlış Pozitif:** Pozitif olarak tahmin edilmiş fakat gerçek değeri negatif olanlar. Türkçe *yp* olarak kısaltılacaktır.
- **Doğru Negatif:** Negatif olarak tahmin edilmiş ve gerçek değeri negatif olanlar. Türkçe *dn* olarak kısaltılacaktır.

Yukarıda Doğru Pozitif (*dp*), Yanlış Negatif (*yn*), Yanlış Pozitif (*yp*), Doğru Negatif (*dn*) değerleri tanımlanmıştı. Bunların oranları hesaplanırken aşağıdaki formüller kullanılmaktadır.

$$dp\ oranı = \frac{dp}{(dp + yn)}$$

$$dn\ oranı = \frac{dn}{(dn + yp)}$$

$$yp\ oranı = \frac{yp}{(yp + dn)}$$

$$yn\ oranı = \frac{yn}{(yn + dp)}$$

Kategorik değişkenleri tahmin ederken yukarıda bahsedilen hesaplama değerlerinin dışında Kesinlik (*Precision*), Hassasiyet (*Recall*) ve F-Ölçüsü (*F-Measure*) değerleri de hesaplanmaktadır. Özellikle sınıfların çok dengesiz olduğu durumlarda yararlanılan bir ölçüttür. Bu değerler aşağıdaki formüller ile hesaplanmaktadır.

$$Hassasiyet\ (p) = \frac{dp}{dp + yn}$$

$$Kesinlik\ (r) = \frac{dp}{dp + yp}$$

$$F - \text{Ölçüsü} = 2 \times \frac{Kesinlik\ \times Hassasiyet}{Kesinlik + Hassasiyet} = 2\frac{pr}{p + r}$$

Uygulama ve Başarımlar

Birçok sınıflandırma yöntemi bulunmaktadır. Weka programı ile bu çalışmadaki C adlı gazetenin verileri farklı yöntemler ile analiz edilmiş ve en iyi başarı yapay sinir ağları ile elde edilmiştir. Bu sebeple sonraki analizlere de yapay sinir ağları devam edilmiştir.

Tablo 1. Farklı Analizler ile Tahmin Sonuçları

Analiz Adı	**F-Ölçüsü**
MultiLayerPerceptron	0,869
NaiveBayesUpdateable	0,836
IBk	0,801
AdaBoostM1	0,849
InputMappedClassifier	0,029
DecisionTable	0,615
J48	0,682

Dört farklı gazete için öncelikle metinlere bakarak gazete adını tahmin etmek amaçlanmıştır. %88 doğruluk oranı ile gazete tahminleri yapılmıştır. Yapılan analize yönelik değerler aşağıdaki gibidir. Her bir gazete C, M, H, S harfleri ile kısaltılmıştır.

Tablo 2. Gazete Adı Tahmin Sonuçları

	dp Oranı	**yp Oranı**	**Hassasiyet**	**Duyarlılık**	**F-Ölçüsü**	**MCC**	**ROC Alanı**	**PRC Alanı**	**Sınıfı**
	0,861	0,050	0,853	0,861	0,857	0,809	0,958	0,918	C
	0,911	0,047	0,868	0,911	0,889	0,851	0,977	0,910	M
	0,850	0,036	0,885	0,850	0,867	0,825	0,965	0,923	H
	0,930	0,017	0,949	0,930	0,939	0,920	0,973	0,949	S
Ağırlıklı Ortalama	0,888	0,037	0,889	0,888	0,888	0,851	0,968	0,925	

Yukarıdaki Tablo 2 incelendiğinde genel olarak gazete tahmin başarı oranlarının birbirine yakın olduğu görülmektedir. Fakat S kodlu gazetenin diğerlerine göre biraz daha yüksek başarı oranı ile tahmin edildiği görülmüştür. Sınıflandırma durumunu görebilmek için ise aşağıdaki matris tablosu incelenebilir.

Tablo 3. Gazete Adı Tahmini Matris Tablosu

Sınıflandırılmış Veriler				
A	**B**	**C**	**D**	
87	8	4	2	A= C
7	92	2	0	B= M
7	5	85	3	C= H
1	1	5	93	D= S

Tablo 3'teki matris tablosu incelendiğinde köşegenler üzerindeki değerlerin daha yüksek olduğu fakat diğer sınıflandırmalarda da bir kısım hatalı sınıflandırmaların yapıldığı görülmektedir. Örneğin gerçekte S kodlu olan gazeteden beş tanesi H kodlu gazete olarak sınıflandırılmıştır.

Buradaki öğrenme işlemi gerçekleştirilirken çok farklı sayıda iç katmanlar ile birçok kez denemeler yapılmış ve en ideal öğrenme sonucu elde edilmeye çalışılmıştır. F-ölçüsü dikkate alınarak yapılan analizler arasından en yüksek değere sahip deneme dikkate alınmıştır. Bunun için üç katmanlı 19x25x18 değerlerine sahip deneme en başarılı olarak görülmüştür. Öğrenme oranı olarak 0.3 ve 500 iterasyon ile denemeler yapılmıştır. Eğitim ve test verilerinin ayrıştırılmasında ise 10 katlı çapraz doğrulama kullanılmıştır.

Buraya kadar verilerden gazetenin tahmin edilmesi amaçlanmıştır. Buradan sonra da bu kez gazetelerin yazarlarını tahmin etme amaçlanmaktadır. İki farklı şekilde çalışma yürütülmüştür. Öncelikle her bir gazete için veriler yapay sinir ağlarına verilmiş ve metnin yazarı tahmin edilmeye çalışılmıştır. Sonrasında ise tüm gazetelerin köşe yazıları birleştirilmiş ve tek dosya halinde verilmiştir. Böylelikle tüm köşe yazıları arasından yazar tahmini yapılmıştır. Aşağıdaki Tablo 4'te her bir gazete için tahmin başarı değerleri gösterilmiştir.

Tablo 4. Her Bir Gazete İçin Tahmin Sonuçları

		dp Oranı	yp Oranı	Hassasiyet	Duyarlılık	F-Ölçüsü	MCC	ROC Alanı	PRC Alanı	Sınıfı
C Gazetesi		0,900	0,011	0,900	0,900	0,900	0,889	0,995	0,950	4
		1,000	0,022	0,846	1,000	0,917	0,910	0,998	0,986	9
		1,000	0,011	0,909	1,000	0,952	0,948	0,999	0,991	11
		0,800	0,011	0,889	0,800	0,842	0,827	0,880	0,772	12
		0,900	0,022	0,818	0,900	0,857	0,842	0,988	0,892	14
		1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	27
		0,900	0,011	0,900	0,900	0,900	0,889	0,995	0,956	31
		0,700	0,022	0,778	0,700	0,737	0,711	0,916	0,817	32
		0,800	0,022	0,800	0,800	0,800	0,778	0,981	0,894	37
		0,700	0,011	0,875	0,700	0,778	0,762	0,907	0,790	39
	Ağ.Ort.	0,871	0,014	0,871	0,871	0,869	0,856	0,966	0,906	
H Gazetesi		0,800	0,033	0,727	0,800	0,762	0,735	0,983	0,805	1
		0,700	0,067	0,538	0,700	0,609	0,565	0,894	0,697	5
		0,500	0,033	0,625	0,500	0,556	0,516	0,970	0,734	7
		0,800	0,011	0,889	0,800	0,842	0,827	0,974	0,904	16
		0,700	0,011	0,875	0,700	0,778	0,762	0,969	0,899	23
		0,700	0,011	0,875	0,700	0,778	0,762	0,922	0,690	25
		0,900	0,011	0,900	0,900	0,900	0,889	0,996	0,957	29
		0,900	0,000	1,000	0,900	0,947	0,943	0,999	0,991	30
		0,800	0,033	0,727	0,800	0,762	0,735	0,979	0,890	36
		0,900	0,044	0,692	0,900	0,783	0,763	0,980	0,876	38
	Ağ.Ort.	0,770	0,026	0,785	0,770	0,772	0,750	0,967	0,844	
M Gazetesi		0,300	0,011	0,750	0,300	0,429	0,443	0,751	0,468	8
		0,900	0,055	0,643	0,900	0,750	0,730	0,990	0,929	13
		0,600	0,022	0,750	0,600	0,667	0,639	0,982	0,876	17
		1,000	0,033	0,786	1,000	0,880	0,872	0,996	0,969	18
		0,600	0,033	0,667	0,600	0,632	0,594	0,962	0,795	20
		0,900	0,000	1,000	0,900	0,947	0,944	0,935	0,914	22
		0,900	0,022	0,818	0,900	0,857	0,842	0,996	0,965	33
		1,000	0,011	0,909	1,000	0,952	0,948	0,998	0,981	34
		0,900	0,022	0,818	0,900	0,857	0,842	0,981	0,895	35
		0,900	0,011	0,900	0,900	0,900	0,889	0,997	0,977	40
	Ağ.Ort.	0,802	0,022	0,804	0,802	0,788	0,775	0,959	0,878	
S Gazetesi		0,900	0,000	1,000	0,900	0,947	0,943	1,000	1,000	2
		0,800	0,033	0,727	0,800	0,762	0,735	0,993	0,952	3
		0,800	0,044	0,667	0,800	0,727	0,698	0,892	0,814	6
		0,900	0,044	0,692	0,900	0,783	0,763	0,990	0,926	10
		0,900	0,000	1,000	0,900	0,947	0,943	0,932	0,914	15
		0,500	0,044	0,556	0,500	0,526	0,478	0,951	0,696	19
		0,800	0,044	0,667	0,800	0,727	0,698	0,986	0,889	21
		0,600	0,000	1,000	0,600	0,750	0,758	0,919	0,840	24
		0,700	0,033	0,700	0,700	0,700	0,667	0,973	0,849	26
		0,600	0,033	0,667	0,600	0,632	0,594	0,928	0,693	28
	Ağ.Ort.	0,750	0,028	0,768	0,750	0,750	0,728	0,956	0,857	

Her bir gazetenin 10 farklı yazarına ait veriler yapay sinir ağlarına öğretildiğinde Tablo 3'teki sonuçlar ortaya çıkmaktadır. Buradaki veriler gazete açısından incelendiğinde %75 ile %86 oranında doğru tahminler elde edildiği görülmektedir. Gazeteler açısından başarı oranındaki bu farklılıkların seçilen yazarların yazarlık stilleri nedeniyle ortaya çıktığı düşünülmektedir. Örneğin S Gazetesi diğerlerine oranla tahmin başarısı düşük iken kendisinin 15 numaralı ve 2 numaralı yazarların tahmininde %94 doğru sınıflandırma

yapılmıştır. Fakat aynı gazetenin 19 numaralı yazarı %52 oranında doğru sınıflandırma değerine sahiptir. Benzer şekilde C Gazetesi %87 ile en yüksek başarılı sınıflandırma oranına sahip gazetedir. Hatta 9, 11 ve 27 numaralı yazarlar %100 başarı ile sınıflandırılmıştır. Fakat 32 numaralı yazarın sınıflandırma başarısı ise %73'te kalmıştır. Bu durum da göstermektedir ki tahmin başarısı yazarın kendisinden doğrudan etkilenmektedir.

Tablo 5. Tüm Köşe Yazıları İçin Tahmin Sonuçları

	dp Oranı	yp Oranı	Hassasiyet	Duyarlılık	F-Ölçüsü	MCC	ROC Alanı	PRC Alanı	Sınıfı
	0,700	0,018	0,500	0,700	0,583	0,579	0,988	0,639	1
	1,000	0,008	0,769	1,000	0,870	0,874	0,998	0,928	2
	0,800	0,008	0,727	0,800	0,762	0,756	0,997	0,914	3
	0,500	0,013	0,500	0,500	0,500	0,487	0,975	0,652	4
	0,500	0,015	0,455	0,500	0,476	0,463	0,942	0,499	5
	0,800	0,008	0,727	0,800	0,762	0,756	0,942	0,845	6
	0,300	0,010	0,429	0,300	0,353	0,345	0,934	0,375	7
	0,100	0,015	0,143	0,100	0,118	0,101	0,790	0,139	8
	0,909	0,015	0,625	0,909	0,741	0,746	0,992	0,779	9
	0,900	0,005	0,818	0,900	0,857	0,854	0,999	0,983	10
	0,800	0,008	0,727	0,800	0,762	0,756	0,997	0,892	11
	0,900	0,003	0,900	0,900	0,900	0,897	0,958	0,896	12
	0,800	0,005	0,800	0,800	0,800	0,795	0,997	0,860	13
	0,800	0,000	1,000	0,800	0,889	0,892	0,994	0,912	14
	0,900	0,003	0,900	0,900	0,900	0,897	0,981	0,902	15
	0,700	0,015	0,538	0,700	0,609	0,603	0,990	0,681	16
	0,600	0,000	1,000	0,600	0,750	0,771	0,970	0,695	17
	1,000	0,005	0,846	1,000	0,917	0,918	1,000	0,984	18
	0,800	0,008	0,727	0,800	0,762	0,756	0,996	0,846	19
	0,500	0,008	0,625	0,500	0,556	0,549	0,947	0,700	20
	0,700	0,008	0,700	0,700	0,700	0,692	0,994	0,835	21
	0,800	0,015	0,571	0,800	0,667	0,667	0,963	0,602	22
	0,700	0,003	0,875	0,700	0,778	0,778	0,945	0,840	23
	0,400	0,003	0,800	0,400	0,533	0,559	0,967	0,717	24
	0,500	0,010	0,556	0,500	0,526	0,516	0,933	0,502	25
	0,700	0,008	0,700	0,700	0,700	0,692	0,995	0,849	26
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	27
	0,500	0,005	0,714	0,500	0,588	0,589	0,993	0,818	28
	1,000	0,005	0,833	1,000	0,909	0,911	0,998	0,945	29
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	30
	0,900	0,005	0,818	0,900	0,857	0,854	0,996	0,843	31
	0,600	0,008	0,667	0,600	0,632	0,624	0,935	0,491	32
	0,900	0,003	0,900	0,900	0,900	0,897	0,998	0,935	33
	1,000	0,003	0,909	1,000	0,952	0,952	0,999	0,970	34
	0,400	0,013	0,444	0,400	0,421	0,408	0,943	0,521	35
	0,600	0,003	0,857	0,600	0,706	0,711	0,976	0,784	36
	0,600	0,010	0,600	0,600	0,600	0,590	0,969	0,688	37
	0,800	0,013	0,615	0,800	0,696	0,693	0,990	0,739	38
	0,600	0,003	0,857	0,600	0,706	0,711	0,929	0,562	39
	0,800	0,003	0,889	0,800	0,842	0,840	0,991	0,905	40
Ağ.Ort.	0,721	0,007	0,727	0,721	0,715	0,713	0,973	0,767	

Yukarıdaki Tablo 5'te ise dört adet gazetenin toplam 40 adet yazarına ait veriler yapay sinir ağlarına öğretilmeye çalışılmıştır. Burada %71 doğru sınıflandırma başarısı elde edilmiştir. Fakat tablodaki veriler dikkatlice incelendiğinde 27 ve 30 numaralı yazarların %100 başarı ile sınıflandırıldığı görülmektedir. Ayrıca 8 numaralı yazarın ise %10 oranında başarı ile sınıflandırıldığı görülmektedir. Bu durum da göstermektedir ki daha önce belirtildiği gibi yazarların yazarlık özelliklerinin etkisi yapay sinir ağları modeli üzerinde doğrudan etkilidir. Bu durumları daha iyi analiz etmeye imkân tanıyan matris tablosu aşağıdaki Tablo 6'da verilmiştir.

Tablo 6. Tüm Köşe Yazıları İçin Matris Tablosu

A	B	C	D	E	F	G	H	İ	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	Aİ	AJ	AK	AL	AM	AN	←SINIFLANDIRILMIŞ
7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	A=1
0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	B=2
0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	C=3
0	1	0	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	D=4
2	0	0	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	E=5
1	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	F=6
0	0	0	0	0	0	3	0	1	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	G=7
0	0	0	2	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	H=8
0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	İ=9
0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	J=10
1	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	K=11
0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	L=12
0	1	0	0	0	0	0	1	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	M=13
0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N=14
0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	O=15
0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	P=16
0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	6	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Q=17
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	R=18
0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S=19
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	T=20
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	7	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	U=21
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	V=22
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	7	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	W=23
0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	X=24
1	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	Y=25
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Z=26
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	AA=27
0	0	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	AB=28
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	AC=29
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	AD=30
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	AE=31
0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	1	0	0	0	AF=32
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	AG=33
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	AH=34
1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	1	0	Aİ=35
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	6	0	0	0	0	AJ=36
0	1	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	AK=37
0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	AL=38
0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	6	0	AM=39
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	8	AN=40

Tablo 5'deki matris tablosu incelendiğinde 8 numaralı yazarın 10 tane yazısından yalnızca bir tanesi doğru olarak yerleştirilmiştir. Diğer dokuz tanesi ise diğer yazarların yazılarına sınıflandırılmıştır. Bu durum bu yazarın yazarlık özelliklerinin ortalama olmasından kaynaklı olabilir. Bunun dışında yedi tane yazarın tüm yazıları %100 başarı oranı ile sınıflandırılmıştır. Buradaki analizler yapılırken eğitim ve test verilerinin ayrıştırılmasında 10 katlı çapraz doğrulama yöntemi kullanılmıştır. Öğrenme oranı olarak 0.3 ve 500 iterasyon ile denemeler yapılmıştır. İç katman sayısı olarak 20 farklı şekilde bir, iki, üç ve dört katmanlı ve

her bir katmanda farklı sayıda sinir bulunan modeller ile denemeler yapılmıştır. En iyi sonuç tek katmanlı ve 35 siniri bulunan modelde elde edilmiştir. Yapay sinir ağlarının dezavantajları arasında probleme uygun ağ yapısının deneme yanılma yolu ile bulunması ve ağı oluşturulma kurallarının bulunmamasıdır. Ayrıca yapay sinir ağlarının davranışları açıklanamamakta ve bu nedenle problem için üretilen çözümün nasıl ve neden üretildiği açıklanamamaktadır (Öztemel, 2012). Çalışma sırasında beklenti çok katmanlı bir yapıda daha iyi sonuçlar üretmesi iken tek katmanlı bir ağın daha iyi sonuç verdiği görülmüştür.

Sonuçlar ve Tartışma

Gazetelerin köşe yazarlarının yazıları sık takip edilince onların yazarı bilinmese dahi bir süre sonra onun kimin tarafından yazıldığı tahmin edilebilir. Bu durum yazarın yazarlık özelliklerini tüm yazılarına yansıtmasından kaynaklanmaktadır. Bu çalışmada dört farklı gazetenin her birinden 10 adet yazar tespit edilip her yazarın 10 adet yazısı kayıt altına alınmıştır. Toplamda 400 adet yazı yapay sinir ağları ile incelenmiş ve gazete adı ile köşe yazarı tahmin edilmeye çalışılmıştır. Gazete adı tahmini yapılırken %93 oranında başarılı sınıflandırmalar elde edilmiştir. Yazar tahminlerinde ise kimi gazetelere ait yazar tahminleri %86 başarı oranı ile tahmin edilirken kimi gazetelere ait yazar tahminleri ise %75 başarı oranı ile tahmin edilmiştir. 400 adet yazı birlikte ele alındığında ise %72 başarı oranı ile tahminler elde edilmiştir. Benzer çalışmalarda farklı başarı oranlarının elde edilmesinde yazarların yazarlık özelliklerini yansıtma durumu etkili olmaktadır. Yapılan çalışmanın metin sınıflandırma işlemlerinde epostaların, resmi yazıların, yazarı bilinmeyen metinlerin gerçek yazarını bulmaya imkân tanıyacağı düşünülmektedir. Bu çalışmanın geliştirilerek kelime dil kökenleri bakımından yazar tahmini yapılabileceği önerilmektedir.

Kaynaklar

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

Amasyalı, M. F., Diri, B., Türkoğlu, F. (2006). Farklı özellik vektörleri ile Türkçe dokümanların yazarlarının belirlenmesi. In The Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'2006).

Aşlıyan, R., Günel, K. (2011). A Comparison of Syllabifying Algorithms for Turkish. Advanced Research in Computer Science, 3(1): 58-78.

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

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

Cavnar, W. B. ve Trenkle, J. M. (1994). N-gram-based text categorization, Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. Information Systems Project Management, Jolyon E. Hallows, AMACOM Pres.

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

Fung, G., Mangasarian, O. (2003). The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization.In Proceedings of the 2003 Conference of Diversity in Computing, pp. 42-46, Atlanta, Georgia, USA.

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

Levent, V. E., Diri, B. (2014). Türkçe Dokümanlarda Yapay SinirAğları İle Yazar Tanıma. Akademik Bilişim'14. Mersin Üniversitesi. 5-7.02.2014.

Malikoğlu, G.P.S.N. (2002). Artificial Intelligence 1, Birsen Yayınevi, İstanbul.

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

Öztemel, E. (2012). Yapay Sinir Ağları, Papatya Yayıncılık, Ankara.

Peng F., Schuurmans D. (2003). Combining Naive Bayes and N-gram Language Models for Test Classification, School of Computer Science, University of Waterloo.

Peng F., Wang S., Schuurmans D. (2003). Language and Task Independent Text Categorization with Simple Language Models, School of Computer Science, University of Waterloo.

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Physical Theraphy, 85(3), 258-268.

Soucy, P., Mineau, G.W. (2001). A Simple K-NN Algorithm For Text Categorization. In Proceedings of The First IEEE International Conference On Data Mining (ICDM_01), pp. 647–648, San Jose, CA.

Stamatatos E., Fakotakis N., Kokkinakis G. (2000). "Automatic Text Categorization in Terms of Genre and Author", Computational Linguistics, pp.471-495.

Tweedie, F., Baayen, H. (1998). How Variable may a Constantbe Measures of Lexical Richness in Perspective, Computers and the Humanities, 32(5):323-352

Prediction of Writers and Newspapers by Artificial Neural Networks in Turkish Corner Manuscripts

Extended abstract

The natural language processing area deals with document classification and verification procedures. If a text is to be identified, the most important feature is undoubtedly the features to be used, and these properties will directly affect success. If several different texts belonging to the author of a text are analyzed and analyzed, it will be seen that these texts have some similarities with each other. These similarities will also allow the author of the text to be guessed. Especially after analyzing and analyzing the corner texts in newspaper texts, the article can be determined more easily. As each author's writing style is reflected in the statistical data, author estimation can be done with certain algorithms to be used on the computer side. K-nearest neighbors, naive bays, support vector machines, J48, random forest methods etc. in text classification studies. different methods have been used. When similar studies were performed on the extracted features to determine the author of a text, we looked at word frequencies, letter numbers, sentence lengths, average syllable numbers, total length of the text, and proportion of words to all words.

In this study, 10 writers were selected randomly from each of four different Turkish newspapers, and a total of 10 corner papers were randomly selected for each author. A total of 400 corner writings were recorded. These texts were recorded in a database table consisting of columns of Author Name, Newspaper, Corner Writing.

All records in the dataset are presented using a .net based program developed by the author using the Zemberek library. Thirty features of texts have been specified through the program. Experiments with different methods have been carried out and back propagation artificial neural networks have produced the most successful results. Among the factors that influence success in predicting the author are the most important selected features. But besides these features, language, selected texts and writers also directly or indirectly influence success. The name of the author and the name of the newspaper were given as model outputs. As input; Number of Sentences, Number of Letter, Average Number of Sentences in Paragraph, Average Word Length, Average Number of words in the sentence, Number of Words, Number of Different Words, Number of Dots, Number of Commas, Number of Characters, Number of Paragraphs, Number of Semicolons, Number of Question Marks, Number of Exclamations, Name Number of Word Type, Special Name Number of Word Type, Number of Adjective Word Type, Verbal Word Count, Number of Pronouns, Conjunction Word Count, Prepositions Type Number of Words, Number of Numeric of Word Types, Number of Time Word Type, Number of Question Word Types, Number of Unknown Word Type, Number of Abbreviation, Number of Double Quotes, Number of Single Quotes, Number of Tires, Number of Parentheses are used.

A 10-fold cross-validation method was used to separate training and test data. In addition, one, two, three and four layers of each layer with different numbers of nerve experiments and tried to reach the best learning model. Each modeled number of 500 iterations and 0.3 learning rate were used.

As a result of the study, different accuracy ratios were obtained for each newspaper. The highest success rate was 86.9% while the lowest success rate was 75.0%. In the newspaper name estimation, success rates between 86% and 93% were obtained. The differentiation of success rates is thought to be influential in the authorship characteristics of each newspaper author. The inner layer numbers were tried out in different forms and it was seen that the model with 35 layers of single layer gave the best result. Disadvantages of artificial neural networks include probabilistic networking by trial and error, and lack of networking rules. Furthermore, the behavior of artificial neural networks cannot be explained, and therefore it is not possible to explain how and why the solution produced for the problem is produced. In the study, it was seen that a single-layer network gave better results while the expectation produced better results in a multi-layer structure.

Keywords: *Machine Learning; Author Estimate; Text Boundary*