Gözetimli Makine Öğrenmesiyle Noktalama ve Etkisiz Kelime Sıklıkları Kullanarak Yazar Tanıma

Araştırma Makalesi/Research Article

Tevfik UYAR[1]*, Kübra KARACAN UYAR[2], Emre YAĞLI[3]

[1]Uluslararası Ticaret Bölümü, İktisadi ve İdari Bilimler Fakültesi, İstanbul Kültür Üniversitesi, İstanbul, Türkiye
[2]Matematik Mühendisliği Bölümü, Fen Edebiyat Fakültesi, İstanbul Teknik Üniversitesi, İstanbul, Türkiye
[3]İngiliz Dilbilimi Bölümü, Edebiyat Fakültesi, Hacettepe Üniversitesi, Ankara, Türkiye
t.uyar@iku.edu.tr, karacank16@itu.edu.tr, yagli@hacettepe.edu.tr
(Geliş/Received:26.09.2019; Kabul/Accepted:31.03.2021)
DOI: 10.17671/gazibtd.623629

Özet— Bu çalışmada köşe yazısı uzunluğundaki yazılarda noktalama ve etkisiz kelime kullanım sıklığı gibi basit özniteliklerin yazar tanımada yeterli olduğu ortaya konmuştur. *Cumhuriyet* gazetesi yazarlarından sıkça köşe yazan 6 adedi seçilerek her birinin çalışmanın başladığı tarihten geriye doğru son 120 köşe yazıları alınmış, her bir yazı için bir takım etkisiz kelime ve noktalama işaretlerinin kullanım sıklıklarına dayanan dokuz adet öznitelik elde edilmiştir. Sekiz gözetimli yapay öğrenme algoritması eğitildikten sonra yazının yazarını tanıma başarısı önişlemsiz ve önişlemden geçirilmiş veri kümelerinde ayrı ayrı ölçülmüş, asgari %82 ve azami %92 olmak üzere yüksek isabetli sonuçlar elde edilmiştir. Ölçeklemenin ve temel bileşen analizinin (PCA) başarıyı anlamlı miktarda değiştirmediği, ancak ölçekleme ve boyut azaltma yöntemi olarak doğrusal ayırtaç çözümlemenin (LDA) birlikte kullanılmasının en yakın komşu (kNN) ve Gaussian Naive Bayes (GNB) algoritmalarının yöntemlerin başarılarında yüksek anlamlı ($p<0.001$), destek vektör makineleri (SVM) algoritmasının başarısında ise anlamlı ($p<0.05$) bir fark yarattığı görülmüştür. Ayrıca karar ağacı temelli rasgele orman algoritmasında (RF) öznitelik önem analizi yapılarak cümle başına ortalama kelime sayısının ve virgül kullanma sıklığının en ayırıcı öznitelikler olduğu tespit edilmiştir.

Anahtar Kelimeler— gözetimli öğrenme, sınıflandırma algoritmaları, yapay öğrenme, yazar tanıma

Columnist Identification with Supervised Machine Learning using Punctuation and Stop Word Frequencies

Abstract— This research asserts that such features as the frequency of stop words and punctuation marks are sufficient for author identification of the texts that are column-long. Six of *Cumhuriyet* columnists who periodically write in the newspaper were selected and 120 columns were collected from each. Nine features based on the frequency of particular stop words and punctuation marks were extracted. Eight supervised machine learning algorithms were trained with extracted feature set. Author identification performance of each algorithm was measured. The effect of dimension reduction and scaling on each algorithm were also examined. Following these procedures, minimum 82% and maximum 92% accuracy were obtained. It is also found that scaling or dimension reduction with principal component analysis (PCA) do not create significant difference alone on accuracy scores, while scaling and linear discriminant analysis significantly increases the validation scores of some of algorithms such as support vector machines ($p<0.05$), Gaussian Naïve Bayes, and k-nearest neighbour ($p<0.001$). Moreover, when feature importance of random forest algorithm is analysed, average word count in a sentence and comma frequency are found as the most important features for detecting the authors.

Keywords— artificial learning, author identification, classification algorithms, supervised learning

1. GİRİŞ (INTRODUCTION)

Son yıllarda donanım ve yazılım alanındaki gelişmeler pek çok farklı kaynaktan elde edilen büyük metin verilerine erişme imkânı vermiş, bu imkân da bu metin verilerinden anlamlı örüntüler elde edebilecek hem dinamik hem de ölçeklenebilir algoritmalara olan ihtiyacı artırmıştır [1]. Yazar tanıma algoritmaları buna örnektir. Bu algoritmaların yegâne hedefi yazarı bilinmeyen bir metnin yazarını (ya da yazarının dili, cinsiyeti gibi bazı demografik özelliklerini) tahminlemektir. Bu metin bir program kodu, bir edebi eser ya da -kriminal amaçlarla- mektup veya e-posta olabileceği gibi [2], akademik bir makale de olabilir [3].

Yazar tanıma işi, makine öğrenmesi perspektifinde ele alındığında bir tek etiketli çok sınıflı metin sınıflama problemidir ve problemin çözümü için yazarın üslubunu temsil edecek özelliklerin elde edilmesi gerekmektedir [4]. Bu özellikler özniteliklere dönüştürülerek yapay öğrenme algoritmaları eğitilir ve böylece yazarı bilinmeyen metnin yazarı tayin edilmeye çalışılır. Literatürde öznitelik olarak kullanılabileceği öne sürülen pek çok özellik vardır. Bunlar kelime dağarcığı zenginliği, işlevsel kelimelerin ve cümle öğelerinin kullanım sıklığı, sabit uzunluktaki karakter kombinasyonlarının ölçümüne dayanan n-gramların sayımıdır [4].

1.1. Bireydil (Idiolect)

Dilbilim literatürüne bakıldığında bireylerin konuşma ya da yazma sırasında kendilerine özgü ayırt edici bir şekilde kullandıkları dil biçemi bireydil kavramıyla karşılanmaktadır [5-7]. Bu görüşten hareketle bireyler, mesajı oluştururken kelime seçimi, dilbilgisi kullanımı ve söyleyiş farklılıkları temelinde zaman içinde geliştirdikleri dilbilimsel parmak izine sahip olur [7]. Bu açıdan bakıldığında, bireydil kavramı yazar tanıma üzerine yapılan çalışmaların kuramsal temelini oluşturmaktadır.

1980'lerden 2000'lere kadar gelen literatür bu "parmak izi"ni daha çok metin türleri üzerinde betimlemiştir. Bu çerçevede belirli metin türlerini (Örn., ders kitapları, bilimsel araştırma metinleri, intihar mektupları, vb.) oluşturan yazarların belirli kelimeleri ve dilbilgisi yapılarını kullandıkları belirtilmektedir [8]–[10]. Bireye ilişkin "parmak izi"ni bulgulamayı amaçlayan çalışmalar ise adli dilbilim (İng. *forensic linguistics*) alanı sınırları içinde son elli yılda kendini göstermektedir. Bu çalışmalarda bir derlemi oluşturan metin içinden bu derleme katkı yapan yazarın bulgulanması amaçlanmaktadır (Örn., [11, 12, 13]).

Hem metin türü hem de birey açısından bakıldığında 'sözcenin ayırt edici özelliği' bir ilke [14], [15] olarak yukarıda bahsedilen çalışma alanlarının temel amacıdır. Bu nedenle metnin konusu ve bağlamı bir değişken olmaksızın bireyin metin üretimi sırasında belirli sözlüksel-dilbilgisel seçimlerde bulunması kaçınılmazdır.

1.2 Yazar Tanıma (Author Identification)

Yazar tanıma, oluşturulmuş bir metnin yazarını tanımaya ve/ya da tahmin etmeye yönelik yapılan çalışmaları ifade eden bir alandır. MacLeod ve Grant [16], yazar tanıma çalışmalarına yönelik geliştirilen yöntemlerin dilbilim ve bilgisayar bilimleri alanlarından verildiğini ifade etmektedir.

Dilbilim literatüründe gerçekleştirilen çalışmalardan Chaski [17], Grant ve Baker [18] ve McMenamin [19] tarafından yapılan çalışmalar alana yöntem açısından katkılarda bulunmuştur. Örneğin Chaski [17], yazar tanıma çözümlemesi için dildeki sözdizimsel birimleri, sözdizimsel açıdan ele alınan noktalama işaretlerini (Örn. farklı cümle türlerini art arda yazarken kullanılan ve sözdizimsel yapıda belirgin durumda bulunan noktalı virgül), cümle karmaşıklığını, kelime zenginliğini, okunabilirliği, yazım, noktalama ve dilbilgisi yanlışlarını ele almış, bu çalışmasının sonucunda sözdizimsel birimlerin ve sözdizimsel açıdan anlamlı olan noktalama işaretlerinin yazarları sınıflandırmada daha elverişli olduğunu bulgulamıştır. Grant ve Baker [18] tarafından gerçekleştirilen çalışmada ise araştırmacılar, ortalama kelime, öbek, cümle ve paragraf uzunluğu, kelime türü sıklığı ve dağılımı, eşdizimli kelimeler (İng. *collocation*) ve içerik analizi üzerinden yazar tanıma çalışmasını gerçekleştirmiş, yazar tanıma çalışmalarında araştırmacıların karşılaşabilecekleri en temel sorunlardan biri olarak örnekleme yöntemini işaret etmiş ve temel bileşen analizinin (İng. *principal component analysis*) yazar tanıma çalışmalarında yöntembilimsel bir katkı sağlayabileceğini ifade etmiştir. Yazar tanıma çalışmalarını dilbilimsel teori ile birleştirme amacıyla McMenamin [19] ise biçembilimsel (İng. *stylistic*) bir yaklaşımla metinleri incelemiş ve hem grup hem de birey açısından dil kullanımının değişmez özelliklerini ortaya koymuştur.

Bilgisayar bilimleri literatürüne bakıldığında Argamon [20], Hoover [21] ve Koppel, Schler ve Argamon [22] tarafından yapılan çalışmalar öne çıkmaktadır. Argamon [22], çalışmasında Burrows [23] tarafından yazar tanımaya yönelik verilen Delta hesaplamaları üzerine yoğunlaşmış ve bu hesaplamaya yönelik kuramsal bir yaklaşım sunmuştur. Hoover [21], çalışmasında metinleri yazara göre sınıflandıran yazar tanıma ile dil kullanım biçemi ve

bu biçemin değişkenlerine odaklanan biçembilimi yöntembilimsel açıdan birleştirmeyi amaçlamış ve çok değişkenli analizle (İng. *multivariate analysis*) bu amacını gerçekleştirmiştir. Koppel ve diğerleri [22] ise sınırlı bir derlem üzerinde yazar tanımayı amaçlayan önceki çalışmalardan farklı olarak geniş bir derlem üzerinde bu işlemin yapılabileceğini göstermiştir. Türkçe literatürde de yazar tanıma amaçlı çalışmalar gerçekleştirilmiştir. Ne var ki Levent ve Diri, ilk örnekleri 1970'lerde özel sözlükler için belirlenmiş sınıf etiketlerine dayanarak doküman tasnifleme olarak başlayan yazar tanıma çalışmalarının tarihsel gelişimine yer verdikleri 2014 tarihli makalelerinde, yazar tanıma uygulamalarının başka dillerde oldukça popüler olmasına karşın Türkçe uygulama sayısının az olduğunu belirtmişlerdir [24]. Türkçe üzerine yapılan yazar tanıma çalışmalarına örnek olarak Levent ve Diri [24], Bozkurt, Bağlıoğlu ve Uyar [25], Taş ve Görür [26], Türkoğlu, Diri ve Amasyalı [27], Doğan ve Diri [28], Yasdi ve Diri [29], Amasyalı, Diri ve Türkoğlu [30], Bay ve Celebi [31], Saygılı, Amghar, Levrat ve Acarman [32] ile Kuyumcu, Buluz ve Kömeçoğlu [33] tarafından yapılan çalışmalar verilebilir.

Tablo 1. Türkçe literatürdeki bazı çalışmaların özellik ve sonuçları
(Properties and results of a number of studies from Turkish literature)

Türkçe literatürdeki bazı çalışmalar	Çalışmaların özellik ve sonuçları		
	Yazar sayısı ve Derlem	*Öznitelikler*	*En yüksek başarı*
Levent ve Diri [24] *	4, 8, 12 ve 16 yazar, 50'şer yazı	16 yazı istatistiği	%95 (ANN)
Bozkurt, Bağlıoğlu ve Uyar [25]	18 yazar (Toplam 500 yazı)	476 işlevsel kelime	%100 (PCA + GPC)
Taş ve Görür [26]	20 yazar 20'şer yazı	35 yazı istatistiği	%80 (CFS** + NBM)
Diri, Amasyalı ve Türkoğlu [27]	9 Yazar, 35'er yazı	(470 ngram)	%96,9 (ANN)
Doğan ve Diri [28]	20 Yazar 40'ar yazı	142 ila 324 (ngram)	%70 (Ng-ind***)
Yasdi ve Diri [29]	10 Yazar 10'ar Yazı	Soyut Özellik Çıkarım (AFE)	%99 (kNN)
Amasyalı, Diri ve Türkoğlu [30]	18 Yazar 35'er Yazı	2148 (Yazı istatistikleri ve ngram)	%92.5 (SVM)
Bay ve Celebi [31]	17 Yazar, 50'şer yazı	20 yazı istatistiği	%99.7 (kNN)
Saygılı, Amghar, Levrat ve Acarman [32]	8 Yazar 50'şer yazı ve 7 yazar 250'şer yazı	Fiilimsi sayıları	(SVM)****
Kuyumcu, Buluz, Kömeçoğlu [33]	237 Yazar, 100'er yazı	100000 (ngram)	%89.6 (Ridge Regresyon)

Notlar: Yazarlar bu çalışmada farklı yazar sayısı ve derlemlere sahip birden fazla sayıda veri setinde yazar tanıma çalışmaları yaptığından birden fazla sonuca ulaşmışlardır. ** CFS: Correlation Feature Selection. *** Ng-ind yazarlar tarafından geliştirilen özgün bir yöntemdir. **** Doğrulama ya da test başarısı yerine her bir yazar için duyarlık ve doğruluk değerleri raporlanmıştır.*

Türkçede yapılan çalışmaların tanımayı amaçladığı yazar kümesinin büyüklüğü, derlemin genişliği, kullanılan özniteliklerin tür ve özellikleriyle elde ettikleri en yüksek başarılara Tablo 1'de özet olarak yer verilmiştir. Bu araştırmalardan bizim çalışmamızla mukayese edilebilir olanların ayrıntılarına sonuç bölümünde yer verilmiştir.

2. YÖNTEM (METHOD)

8 Ocak 2019 tarihinde Cumhuriyet Gazetesi internet sitesinin tüm yazarlar sayfasında en üstte yer alan altı adet yazar (Ali Sirmen, Erdal Atabek, Ergin Yıldızoğlu, Mine Söğüt, Mustafa Balbay ve Şükran Soner) yazar kümesi olarak belirlenmiştir. Her bir yazarın 8 Ocak 2019 tarihinden geriye olmak üzere son 120'şer yazısı ayrı metin dosyaları şeklinde kaydedilmiştir. Toplamda 720 adet yazı için öznitelikleri elde etmeden önce, paragraf ayrımlarında kullanılan "***" ve "* * *" işaretleri temizlenmiş, tüm noktalama işaretleri saydırılıp kullanacağımız öz niteliklerle ilişkili olanların sayıları kaydedildikten sonra yazılar noktalama işaretlerinden temizlenmiştir.

Her bir nokta, soru işareti, ünlem ve üç nokta işaretinin cümle sonunu temsil ettiği varsayılarak cümle sayıları elde edildikten sonra aşağıdaki öznitelikler hesaplanmıştır:

- Yazıdaki "ve" bağlacı sıklığı,
- Yazıdaki "bir" kelimesi sıklığı,
- Yazıdaki "bu", "şu", "o" zamir ve sıfatları sıklığı,
- Yazıdaki "de", "da" bağlacı sıklığı,
- Virgül kullanma sıklığı,
- Noktalı virgül kullanma sıklığı,
- Tırnak işareti kullanma sıklığı,
- Cümle sonlarında nokta dışında işaret kullanım sıklığı
- Cümle başına ortalama kelime sayısı

Sıklık ifadesi, ilgili kelime/işaret sayısının cümle sayısına bölünmesiyle elde edilen değer anlamına gelmektedir. Bu sayede yazıların kısalık ve uzunluklarından kaynaklanan farklar bertaraf edilmiş ve standardizasyon sağlanmıştır. Tırnak işaretinin kesme işaretinden ayırt edildiğinden emin olunmuş ve alıntı için kullanılan çift tırnak işareti (' " ') miktarı sayılmıştır. Kelimelerin olası kullanım amacı farklılıkları ihmal edilmiştir (Örneğin yazıdaki "de", "da" kelimelerinin her birinin bağlaç olduğu varsayılmıştır).

Bu yolla 720 x 9 boyutlarındaki veri kümesi oluşturulmuştur. Bu veri kümesinin %20'si test kümesi olarak ayrıldıktan sonra kalan 576 verilik eğitim setiyle

Python *scikitlearn* kütüphanesinde tanımlı olan aşağıdaki algoritmalar eğitilmiştir:

- Lojistik Bağlanım (*Logistic Regression,* LR),
- Doğrusal Ayırtaç Çözümlemesi (*Linear Discriminant Analysis,* LDA),
- Kuadratik Ayırtaç Çözümlemesi (*Quadratic Discriminant Analysis,* QDA),
- En Yakın Komşular (*k-Nearest Neighbors,* kNN),
- Gaussian Naive Bayes (GNB),
- Rasgele Orman (*Random Forest,* RF),
- Gradient Boosting (GBC) ve
- Yapay Sinir Ağları (*Artificial Neural Networks,* ANN)

2.1 Sınıflama Algoritmaları (Classification Alghorithms)

Bu çalışmada kullandığımız sınıflama algoritmaları ve bazı özellikleri aşağıda açıklanmıştır:

Lojistik Bağlanım: Sınıflandırma problemlerinde sıkça kullanılan gözetimli makine öğrenmesi yöntemlerindendir. Bu yöntemde verilen her bir x_i özniteliği için, verinin istenen etiketlerde olma olasılıkları tek tek hesaplanır ve elde edilen tüm olasılıklar çarpılır. Bu çarpım, Olabilirlik Fonksiyonu olarak adlandırılır. Bu algoritmanın amacı Olabilirlik Fonksiyonunu maksimize ederek yeni parametre değerlerini bulmaktır. Bu yöntemde (0,1) tanım aralığında, iki etiketin tek bir öznitelikle olan benzerliği ait (sınıflandırma) bir lojistik fonksiyon kullanılarak elde edilir [34].

Gaussian Naive Bayes: Bayes sınıflandırıcıları, Bayes teoremini her öznitelik çifti arasındaki 'saf' bağımsızlık varsayımıyla uygulayan bir dizi gözetimli makine öğrenmesi algoritmasıdır. Bu Naive Bayes sınıflandırıcısı ise verilen örneğin belirli bir sınıfa ait olma olasılığını Gauss Dağılımına uygun olduğu varsayımıyla hesaplar [35].

Destek Vektör Makineleri: İstatistiksel öğrenme teorisine dayalı gözetimli makine öğrenmesi tekniklerinden biri de destek vektör makineleri yöntemidir. Elimizde p-boyutlu bir öznitelik matrisi ve veri kümesine ait sınıf bilgileri olsun. Burada yöntemin temel amacı, iki veya daha çok sınıfa ait verileri bir düzlem (hyperplane) yardımıyla en uygun şekilde iki sınıfa ayırmaktır.

$$Düzlem: \{X \in \mathbb{R}^p : \beta^T X + \beta_0 = 0, \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}\}$$

Denklemde sırasıyla β^T ve β_0 değeri eğim ve kesme noktasına karşılık gelmektedir. Ayırmak istenilen iki sınıf arasından geçecek en uygun düzlem belirlenerek sınıflandırma yapılır [34].

Doğrusal ve Kuadratik Ayırtaç Çözümlemeleri (LDA & QDA): Ayırtaç çözümlemeleri yüksek boyutlu veri analizlerinde, sıklıkla kullanılan gözetimli boyut azaltma yöntemleridir. Bu yöntemlerde ilk olarak bir alt küme seçimi yapılır. Bu seçim, farklı etiketlere ait veri kümlerini birbirinden en iyi biçimde ayırmak için yapılır. Seçilen alt küme ile birlikte daha az sayıda fakat veri kümesini en iyi biçimde temsil edebilecek özniteliklere ulaşılmış olur [36].

LDA yönteminde farklı sınıflara ait ortalamaların arasındaki, daha sonra bir sınıfın ortalaması ile sınıfı bulanacak örnek arasındaki uzaklıklar hesaplanır [37].

LDA'de öznitelik kovaryans matrislerinin eşit olduğu varsayılırken, QDA her bir sınıf için farklı kovaryans matrislerine müsaade eder ve bu nedenle de karar ayırtacı doğrusal olmak durumunda değildir.

Gradient Boosting (GBC) ve Rastgele Orman (RF): Ağaç sınıflandırmalarının topluluk (İng. *ensemble*) yöntemleridir. Gradient Boosting bir takviye (İng. *boosting*), rasgele orman ise bir torbalama (İng. *bagging*) yöntemidir [38].

Gradient Boosting'de iteratif olarak her birisi öncekindeki hata miktarını azaltan yeni bir ağaç oluşturma yoluyla çalışırken, Rasgele ormanda veriden elde edilen rasgele altkümelerin her biri için bağımsız ağaçlar oluşturulur, her biri ayrı ayrı eğitilir ve ağaç içindeki dallanmalar korelasyonu azaltacak biçimde seçilir [38, 39].

k-En Yakın Komşular (kNN): İstatistiksel öğrenme teorisine dayalı örnek tabanlı gözetimli makine öğrenmesi yöntemidir. Sınıfları önceden belirlenmiş bir veri kümesi yardımıyla sınıflandırılmamış bir verinin etiketi, belirlenen k en yakın komşunun etiketine bakarak belirlenir. Böylece, sınıfı belirlenmek istenen verinin etiketi, k adet komşu incelenip en sık kullanılan etikete tanımlanır [38]. Bu çalışmada k=13 en yakın komşu belirlenmiş ve top ağacı algoritması kullanılmıştır.

Yapay Sinir Ağları (ANN): Biyolojik sinir ağlarından ilham alan yapay sinir ağları, yüksek tahmin ve modelleme yeteneğiyle makine öğrenmesi algoritmaları içinde önemli bir yere sahiptir. Yapay sinir ağları çok karmaşık veya tanımsız bir işlevi temsil eden bir girdi-çıktı veri kümesi için eğitilirler. Çok sayıda gizli katman ve nöronla, herhangi bir girdi-çıktı ilişkisini modelleyebilir [40].

Bu çalışmada küçük veri kümelerinde başarılı olan *relu* aktivasyonu ve *lbfgs* çözücüsü kullanılmıştır. Öğrenme oranı sabit tutulmuş, iki gizli katmanda 50'şer düğüm tercih edilmiştir.

2.2 Önişlemler (Preprocessing)

Veri önişleme, eğitim setinde ekleme, çıkarma ve dönüştürme işlemlerine verilen genel addır. Verinin

dönüştürülmesi, öğrenme algoritmasının performansını önemli ölçüde etkileme potansiyeline sahiptir [39].

Çalışmamızda algoritmaların veri kümesindeki doğrudan isabet başarısını ölçmenin yanı sıra, başarıyı artırıp artırmadığını değerlendirmek amacıyla birtakım önişlemlere de başvurulmuştur. İncelediğimiz bu önişlemler ölçekleme, bileşen azaltma ve boyut azaltmadır.

Ölçekleme, özellikle farkı ölçeklere sahip öznitelikler bulunduğunda bu öz nitelikleri aynı aralıkta normalize etmek için kullanılır. Temel bileşen çözümlemesi (Principal Component Analysis, PCA) ise, verideki değişim miktarlarını ölçüt olarak sayıp veri noktaları arasındaki farkın en iyi olduğu öznitelikleri bulur ve temel bileşen olarak kabul eder. Sınıf bilgisi kullanmadığından aslında daha çok gözetimsiz makine öğrenme tekniklerinde kullanılır [36].

Çalışmamızda boyut azaltma yöntemi olarak doğrusal ayırtaç çözümlemesi (LDA) kullanılmıştır. Daha önce de belirtildiği üzere LDA gözetimli bir boyut azaltma yöntemi olup, alt veri kümelerinden veriyi en iyi temsil eden kümenin eldesine dayanır.

3. BULGULAR (RESULTS)

Çalışmamız CORE İ7 7700HQ 2.8GHZ işlemci ile Python programlama dili ve *scikitlearn* kütüphanesi ile gerçekleştirilmiştir. Yöntemler, (i) hiçbir önişlemden geçirilmeyen ham veri, (ii) ölçeklenmiş veri, (iii) temel bileşen analiziyle iki bileşene indirgenmiş veri, (iv) ölçeklenmiş ve LDA ile iyi boyuta indirgenmiş veri kümelerinin her biriyle sınanmıştır.

Önişlemden geçirilmemiş ham veriden ve hem ölçeklenmiş hem de LDA ile iki boyuta indirgenmiş önişlemli veriden 30 katlı çapraz geçerlilik yöntemiyle elde edilen ortalama doğrulama skorları ve standart sapmaları Tablo 2'de gösterilmiştir.

Tablo 2. Algoritmaların ham ve ölçeklenmiş + boyutu azaltılmış veriyle performansları
(Validation and test scores of the algorithms with raw and scaled+dimensionally reduced data)

Alg.	İsabet Skoru				
	Doğrulama	*Test*	*Ölçek + LDA*	*Ölçek + LDA Test*	*p-değeri*
LR	0,86 ± 0,08	0,88	0,87 ± 0,07	0,90	
LDA	0,83 ± 0,08	0,90	0,83 ± 0,10	0,90	
QDA	0,85 ± 0,09	0,92	0,88 ± 0,08	0,90	
kNN	0,80 ± 0,08	0,82	0,88 ± 0,07	0,92	p < 0,001
GNB	0,80 ± 0,08	0,86	0,87 ± 0,08	0,90	p < 0,001
SVM	0,86 ± 0,08	0,88	0,90 ± 0,06	0,92	p < 0,05
RF	0,84 ± 0,08	0,88	0,85 ± 0,08	0,85	
GBC	0,87 ± 0,07	0,90	0,89 ± 0,08	0,89	
ANN	0,86 ± 0,06	0,88	0,86 ± 0,07	0,86	

Önişlemenin anlamlı bir fark yaratıp yaratmadığını tespit etmede 30 katlı çarpaz geçerlilik sonuçların kullanılmış ve t-testi uygulanmıştır.

Veri kümesi sadece ölçeklendiğinde ve temel bileşen analiziyle indirgendiğinde isabet başarısında anlamlı hiçbir fark oluşmadığından bu sınamaların sonuçlarına Tablo 2'de yer verilmemiştir.

Tablo 2'de de görüleceği üzere, ölçeklemek ve boyut azaltmak en yakın komşu (kNN) ve Gausian Naive Bayes (GNB) yöntemlerinin doğrulama başarısında yüksek derecede anlamlı (p<0,001), destek vektör makinelerinde (SVM) ise anlamlı bir fark yaratmıştır.

En başarılı test sonucu %92 olmak suretiyle kNN ve SVM ile elde edilmiştir. Beklendiği üzere ölçekleme ve LDA ie boyut azaltma zaten boyut azaltmanın yöntemin kendisi olduğu LDA ve QDA yöntemlerinde bir fark yaratmamıştır.

Tablo 3. Rasgele orman yönteminde öznitelik önem sırası
(Feature importance order of random forest method)

RF	Öznitelikler ve Önemleri	
	Öznitelik	*Yüzde*
1	Cümle başına ortalama kelime sayısı	%24,76
2	Virgül kullanma sıklığı	%21,86
3	"ve" bağlacı sıklığı	%14,58
4	"de", "da" bağlacı sıklığı	%7,69
5	Tırnak işareti kullanma sıklığı	%7,02
6	Cümle sonlarında nokta dışında işaret kullanım sıklığı	%7,01
7	"bir" kelimesi sıklığı	%6,19
8	Noktalı virgül kullanma sıklığı	%5,67
9	"bu", "şu", "o" zamir ve sıfatları sıklığı	%5,22
TOPLAM		%100

Rasgele orman yöntemi analiz edilerek başarıda etkili öznitelikler önemine göre sıralandığında Tablo 3'te görülen sıralama elde edilmiştir. Geliştirilmiş bir karar ağacı yöntemi olan rasgele orman algoritmasına göre, en ayırıcı öznitelikler sırasıyla cümle başına kelime sayısı (%24,76), virgül sıklığı (%21,86) ve "ve bağlacı sıklığı" (%14,58) olarak öne çıkmıştır. Diğer öznitelikler %5 ila %7 arasında birbirinden çok da farklı olmayan önemde görünmektedirler.

4. SONUÇ (CONCLUSION)

Bu çalışmada kullanılan yöntemlerle 6 köşe yazarının metinleri etkisiz kelime ve bazı yazı özellikler kullanılarak %92 başarıya ulaşılmıştır. Sadece 9 basit öznitelik kullanıldığı göz önünde bulundurulursa, yazar tanımada yüksek isabetin zaman ve hesaplama karmaşıklığı bakımından oldukça tutumlu yöntemlerle de elde edilebileceği anlaşılmıştır.

Daha önce Türkçe literatürde yapılan çalışmalarda ağırlıklı olarak n-gram yönteminin kullanıldığı, n-gramlardan büyük boyutlu özellik vektörleri oluşturularak sınıflama yapıldığı görülmektedir. Bu tarz yöntemler öznitelik sayısı fazlalığından ötürü tutumlu değillerdir, dolayısıyla uzun zaman ve yüksek hesaplama gücü gerektirirler. Örneğin Diri, Amasyalı ve Türkoğlu [27], 40 adet yazı istatistiğine dayalı özniteliğin yanısıra n-gramları da içeren farklı veri setlerini kullandıkları çalışmalarında, azami başarılarını 470 elemanlı 2-gram öznitelikleri içeren veri setiyle, yapay sinir ağları kullanarak 9 yazarı ayırt etmede %96,9 olarak kaydetmişlerdir ancak hem öznitelik sayısı hem de kullanılan yöntemin karmaşıklığı sebebiyle -yazarlar hesaplama süresini belirtmemiş olsalar da- oldukça yüksek hesaplama süresi gerektirdiğini tahmin etmekteyiz. Destek vektör makineleri kullanarak bizimle aynı isabet oranını kaydeden Amasyalı, Diri ve Türkoğlu [30] çalışmasına bakıldığında da 2-gram, 3-gram ve çeşitli yazı istatistiklerini içeren 2148 öznitelikle çalışıldığı görülmüştür. Yine n-gram'lerin eldesinin ve öznitelik vektörünün uzunluğu nedeniyle eğitim ve test sürelerinin gerektirdiği hesaplama karmaşıklığı hesaba katıldığında bizim çalışmamızda kullanılan yöntem ile aynı sonuca çok daha basit bir şekilde ulaşıldığı güvenli bir şekilde söylenebilir.

Kullanılan veri kümeleri farklı olduğundan literatürdeki diğer çalışmalarla birebir mukayese imkânı olmamakla birlikte, taradığımız Türkçe literatürde bizim çalışmamıza benzer şekilde sadece yazı istatistiklerine dayalı tanıma gerçekleştirilen çalışmalar, Levent ve Diri [24], Taş ve Görür [26] ile Bay ve Celebi [38] tarafından yapılan çalışmalardır.

Çalışmamız Levent ve Diri [24] ile karşılaştırıldığında, söz konusu çalışmada 4 yazarlı veri setinde %100, 8 yazarlı veri setinde %78, bu iki veri setinin birleştirildiği 12 yazarlı veri setindeyse %95 olduğu görülmektedir. Yazarlar 8 yazarlı veri setinde elde edilen isabetin 12 yazarlı veri setinden daha düşük (%78) olmasını dört yazarın yazarlık özelliklerinin çok daha iyi ayırt edilmesiyle açıklamıştır. Ancak 16 öznitelik ve üç katmanlı bir yapay sinir ağı kullandıklarından, kendilerinin de çalışmalarında ifade ettikleri üzere, bu başarılı sonucun elde edilmesinin oldukça zaman aldığı görülmektedir. Örneğin %95'lik başarı AMD A8-5550M APU 2.10 GHz işlemci, 4 GB hafıza, 64 bit işletim sistemi Windows 8.1 işletim sisteminde .Net ortamında geliştirilen uygulamayla 44,37 dakikada elde edilmiştir. Daha az öz nitelikle ve yapay sinir ağlarına nispetle daha az karmaşıklığa sahip algoritmalara başvurduğumuz çalışmamızda hesaplama süreleri çok daha düşüktür. Sözgelimi %92 test başarısı elde edilen SVM algoritmasının tüm ön işlemler, eğitim, doğrulama ve test toplam sonuç verme süresi 21,24 saniye olarak ölçülmüştür.

20 yazardan 20'şer yazının kullanıldığı Taş ve Görür [26] çalışmasında 35 öznitelik kullanılmış, 15 algoritmanın başarısı ölçülmüş ve kullanılan algoritmalar arasında en yüksek isabet oranı %70.75 olarak kaydedilmiştir. Daha sonra öznitelik analizi yapılmış ve öz nitelik sayısı 22'ye düşürüldükten sonra azami %80'lik bir isabet oranı elde edilebilmiştir. Yazar sayısı ve her bir yazar için kullanılan yazı sayısına bakıldığında %80 oldukça yüksek bir performans olarak değerlendirilebilir. Ne var ki tanılayıcı öznitelik sayısının oldukça yüksektir ve diğer çalışmalarda da olduğu gibi bu durum hesaplama karmaşıklığını ciddi ölçüde artırmaktadır.

Bay ve Celebi [38], 7 yazardan 50'şer yazı kullandıkları ve 20 öznitelik elde ettikleri veri setinde kNN yöntemiyle %99.71 doğrulama başarısı elde etmişlerdir. Hatta ki-kare yöntemiyle öznitelik elemesi yapmışlar ve 17 öznitelikle aynı veri setinde %100 doğrulama skoru kaydetmişlerdir. Kullanılan algoritmaların niteliği ve büyük ölçüde noktalama sayısı ve noktalama sıklığına dayanan öznitelikleri bakımından bizim çalışmamıza en çok benzerlik gösteren çalışma budur ve oldukça yüksek bir başarı elde edilmiştir. Bu çalışmada bizim çalışmamızın tam tersine, öznitelikler elde edilmeden önce etkisiz kelimeler metinlerden tamamen çıkarılmıştır. Ancak yazarlar sadece 10 katlı çapraz geçerlilik ortalama doğrulama skorlarını raporlamışlardır. Yöntemin test başarısı raporlanmadığından böylesine yüksek bir isabet oranının aşırı öğrenmeye sebebiyet verip vermediği, algoritmaların eğitim ve doğrulama testleri sırasında hiç görmediği verilerde ne kadar başarı sağlayacağı anlaşılamamaktadır.

Çalışmamızda özniteliklerin basitliği ve azlığına karşın öne sürdüğümüz yöntemin genel olarak başarılı bir sonuç verdiğini söylemek mümkündür. Ancak yazar sayısının altıyla sınırlı tutulmuş olması araştırmamızın en önemli kısıtını oluşturmaktadır. Yazar sayısı arttıkça bu başarının düşeceği de aşikârdır. Ayrıca gazeteye gönderilen köşe yazılarının editör müdahalesine ne kadar maruz kaldığı, yazarların yazılarının farklı editörlerce düzeltilip düzeltilmediği bilinmemektedir ve araştırma bu belirsizlik altında gerçekleştirilmiştir.

Yöntemin sadece yazı istatistiklerine dayanması sebebiyle yazı uzunluğunun yöntem başarısı üzerinde belirleyici olduğu, daha kısa metinlerde aynı başarının elde edilemeyeceği, buna mukabil, daha uzun metinlerde bireydilin daha çok ön plana çıkacağı varsayımıyla daha başarılı olabileceğini düşünmekteyiz. Bu bakımdan kullandığımız yöntemin daha uzun metinlerde yazar tanıma için, editör müdahalesinin kayda değer olduğu kitap ölçeğindeki metinlerde de editör tanıma için kullanılabileceğini düşünmekteyiz.

KAYNAKLAR (REFERENCES)

[1] C. C. Aggarwal, C. X. Zhai, "An introduction to text mining", **Mining Text Data**, Editör: Aggarwal, C. C., Zhai, C. X., Springer, Boston, MA, A.B.D., 1–10, 2013.

[2] O. de Vel, A. Anderson, M. Corney, G. Mohay, "Mining e-mail content for author identification forensics", *ACM SIGMOD Record*, 30(4), 55-64, Ara. 2001.

[3] S. Hill ve F. Provost, "The myth of the double-blind review?", *ACM SIGKDD Explorations Newsletter*, 5(2), 179-184, 2003.

[4] J. Houvardas ve E. Stamatatos, "N-Gram Feature Selection for Authorship Identification", **Artificial Intelligence: Methodology, Systems, and Applications. AIMSA 2006,** Cilt 4183, Editör: Euzenat J., Domingue J.. Springer, Berlin, Heidelberg, 77-86, 2006.

[5] D. Abercrombie, "Voice qualities", **Psycholinguistics: An introduction to the study of speech and personality**, Editör: Markel, N.N., The Dorsey Press, Londra, 109–127, 1969.

[6] M. A. K. Halliday, A. McIntosh, ve P. Strevens, **The linguistic sciences and language teaching**, Longman, Londra, 1964.

[7] M. Coulthard, "Author identification, idiolect, and linguistic uniqueness", *Appl. Linguist*., 25(4),. 431–447, 2004.

[8] D. Biber, **Variation across speech and writing,** Cambridge University Press, Cambridge, 1988.

[9] D. Biber, **Dimensions of register variation: A cross-linguistic comparison.** Cambridge University Press, Cambridge, 1995.

[10] R. Shuy, **The language of confession, interrogation and deception,** Sage, Londra, 1998.

[11] M. Coulthard, "Forensic discourse analysis", **Advances in spoken discourse analysis,** Editör: Coulthard, N. Routledge, Londra, 242–257, 1992.

[12] M. Coulthard, "On the use of corpora in the analysis of forensic texts", *Forensic Linguist. Int. J. Speech, Lang. Law*, 1(1), 27–43, 1994.

[13] R. Eagleson, "Forensic analysis of personal written text: A case study", **Language and the law**, Editör: Gibbons, J., Longman, Londra, 362–373, 1994.

[14] N. Chomsky, **Aspects of the theory of syntax,** MIT Press, Cambridge, 1965.

[15] M. A. K. Halliday, **Learning how to mean,** Edward Arnold, Londra, 1975.

[16] N. MacLeod, T. Grant, "Whose Tweet? Authorship analysis of micro-blogs and other short-form messages", **International Association of Forensic Linguists' Tenth Biennial Conference,** 210–224, 2012.

[17] C. Chaski, "Empirical evaluations of language-based authorship identification techniques", *Int. J. Speech,* Lang. Law, 8(1), 1–65, 2001.

[18] T. Grant ve K. Baker, "Identifying reliable, valid markers of authorship: A response to Chaski", *Int. J. Speech, Lang. Law*, 8(1), 66–79, 2001.

[19] G. R. McMenamin, "Style markers in authortship studies", *Int. J. Speech, Lang. Law*, 8(2), 93–97, 2001.

[20] S. Argamon, "Interpreting Burrows's Delta: geometric and probabilistic foundations", *Lit. Linguist. Comput.,* 23(2), 131–147, 2008.

[21] D. L. Hoover, "Multivariate analysis and the study of style variation", *Lit. Linguist. Comput.,* 18(4), 341–359, 2003.

[22] M. Koppel, J. Schler, ve S. Argamon, "Authorship attribution in the wild", *Lang. Resour. Eval.,* 45, 83–94, 2011.

[23] J. Burrows, "Delta: A measure for stylistic difference and a guide to likely authorship*", Lit. Linguist. Comput.,* 17(3), 267–287, 2002.

[24] B. Levent, V. E. Diri, "Türkçe dokümanlarda yapay sinir ağları ile yazar tanıma", **XVI. Akademik Bilişim Konferansı Mersin Üniversitesi,** 735–741, 5 - 7 Şubat 2014.

[25] I. N. Bozkurt, Ö. Bağlıoğlu, ve E. Uyar, "Authorship attribution: performance of various features and classification methods", **22nd International Symposium on Computer and Information Sciences,** ISCIS 2007 - Proceedings, 158–162, 2007.

[26] T. Taş ve A. K. Görür, "Author identification for Turkish texts*", J. Arts Sci.,* 7, 151–161, 2007.

[27] F. Türkoğlu, B. Diri, ve M. F. Amasyalı, "Author attribution of Turkish texts by feature mining", **Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues**, 1086–1093, 2007.

[28] S. Doğan ve B. Diri, "Türkçe dökümanlar için N-gram tabanlı yeni bir sınıflandırma(Ng-ind): Yazar, tür ve cinsiyet", *Türkiye Bilişim Vakfı Bilgi. Bilim. ve Mühendisliği Derg*., 1(3), 11–19, 2010.

[29] M. Yasdi, B. Diri, "Soyut özetllik çıkarımı ile yazar tanıma", **IEEE 20. Sinyal İşleme ve İletişim Uygulamaları Kurultayı,** Fethiye, Muğla, Türkiye, 2012.

[30] M. F. Amasyalı, B. Diri, F. Türkoğlu, "Farklı özellik vektörleri ile Türkçe dökümanların yazarlarının belirlenmesi", **15. Türkiye Yapay Sinir Ağları Sempozyumu**, Muğla, 21- 24 Haziran, 2006.

[31] Y. Bay, E. Çelebi, "Feature Selection for Enhanced Author Identification of Turkish Text", **30th International Symposium on Computer and Information Sciences,** ISCIS 2015 - Proceedings, 371-379, 2015.

[32] N. Ş. Saygılı, T. Amghar, B. Levrat, T. Acarman, "Taking advantage of Turkish characteristic features to achieve authorship attribution problems for Turkish", **25th Signal Processing and Communications Applications Conference (SIU),** Antalya, 2017.

[33] B. Kuyumcu, B. Buluz, Y. Kömeçoğlu, “Author Identification in Turkish Documents with Ridge Regression Analysis”, **27th Signal Processing and Communications Applications Conference (SIU),** Sivas, 24-26 Nisan 2019.

[34] G. James, D. Witten, T. Hastie, R. Tibshirani, **An Introduction to Statistical Learning with Application in R**, Springer, Los Angeles, A.B.D., 2017.

[35] S. B.Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", *Informatica*, 31, 249–268, 2007.

[36] E. Alpaydın, **Yapay Öğrenme**, Boğaziçi Üniversitesi Yayınları, İstanbul, 88-116, 2017.

[37] H.Wang, C. Ding, H. Huang, "Multi-label linear discriminant analysis", **Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),** *6316 LNCS(PART 6)*, 126–139, 2017.

[38] T. Hastie, J. Tibshirani, J. Friedman, **The Elements of Statistical Learning, Data Mining, Inference, and Prediction,** Springer, New York, A.B.D., 2016.

[39] M.Kuhn, K. Johanson, **Applied Predictive Modeling**, Springer, New York, 2013.

[40] A. G. Karacor, E.Torun, R. Abay, “Aircraft Classification Using Image Processing Tecniques and Artificial Neural Neworks", *International Journal of Pattern Recognition and Artificial Intelligence*, 25(08), 1321–1335. 2011.