Fırat Üniversitesi Müh. Bil. Dergisi
33(2), 519-526, 2021

Araştırma Makalesi
https://doi.org/10.35234/fumbd.871986

Türkçe Haber Metinlerinin Çok Terimli Naive Bayes Algoritması Kullanılarak Sınıflandırılması

Emrah AYDEMİR[1*], Murat IŞIK[2], Türker TUNCER[3]
[1] Yönetim Bilişim Sistemleri Bölümü, İşletme Fakültesi, Sakarya Üniversitesi, Sakarya, Türkiye
[2] Bilgisayar Mühendisliği Bölümü, Mühendislik-Mimarlık Fakültesi, Kırşehir Ahi Evran Üniversitesi, Kırşehir, Türkiye
[2] Adli Bilişim Sistemleri Bölümü, Teknoloji Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye
[*1] emrahaydemir@sakarya.edu.tr, [2] muratisik@ahievran.edu.tr, [3] turkertuncer@firat.edu.tr

(Geliş/Received: 31/01/2021; Kabul/Accepted: 23/04/2021)

Öz: Hızla gelişen teknoloji ile verilere erişmek oldukça kolaylaşmış ancak elde edilen bu veri yığınlarının işlenmesi ve analiz edilmesi ise büyük bir problem haline gelmiştir. Bu çalışmada çevrimiçi bir haber sitesinden metin halinde toplanan yazıların, metin madenciliği ile daha önceden belirlenmiş haber kategorilerine ayrılması sağlanmıştır. Metin halinde toplanan 2248 haber verisi için iki ayrı yöntem kullanılmış ve haberlerin, birinci yöntemde %95,24'ü ikinci yöntemde ise %99,86'sı doğru olarak sınıflandırılmıştır. Türkçe dilinin özgün yapısından kaynaklı sınıflandırma yapılmasının zorluğundan dolayı bu çalışma ileriki metin madenciliği uygulamaları için faydalı olacaktır. Ayrıca elde edilen sonuçlar, literatürde yer edinmiş benzer çalışmalar ile karşılaştırılarak analiz edilmiştir.

Anahtar kelimeler: Metin madenciliği, veri madenciliği, metin sınıflandırma, naive bayes, rastgele orman, makine öğrenmesi.

Classification of Turkish News Texts with Multinomial Naive Bayes Algorithm

Abstract: Rapidly developing technology, it has become quite easy to access data, however the processing and analysis of these collected data have become a major problem. In this study, the news articles collected from an online news website in text form are classified into predefined categories with text mining. Two different methods were applied to 2248 news collected in text-form. The news articles were classified with 95.24% accuracy by applying the first method and 99.86% accuracy by applying the second method. This study will be useful for future text mining applications due to the difficulty of text-classification because of original structure of the Turkish language. In addition, the results were analyzed by comparing them with the similar studies in the related literature.

Key words: Text mining, data mining, text classification, naive bayes, random forest, machine learning.

1. Giriş

Gelişen teknoloji ile günümüzde veri toplama araçları oldukça gelişmiş ve bu sayede veriye ulaşmak oldukça kolaylaşmıştır. Ancak elde edilen verilerin çokluğu; bu verilerin anlamlaştırılması, işlenmesi ve analiz edilmesi gibi birtakım problemleri ortaya çıkarmıştır. Bu problemin çözümünde kullanılan yöntemlerden en önemlisi veri madenciliği tekniğidir [1]. En basit anlamıyla veri madenciliği, toplanan veri yığınından anlamlı bir bilgi çıkarma işlemidir. Günümüzde veri madenciliğinin; pazarlama yönetimi, bankacılık, sınıflandırma ve gruplama, maliye, finans, borsa, satış yönetimi, sigortacılık, telekomünikasyon, elektronik ticaret, sağlık sistemleri, tıp, biyoloji, biyokimya, genetik, endüstriyel analiz ve çözümler, eğitim, istihbarat, bilim ve mühendislik [2–5] gibi oldukça geniş kullanım alanları mevcuttur. Veri yığınının türüne ve kullanım amacına göre uygulanacak veri madenciliği tekniği değişmektedir. Metin madenciliği bu tekniklerden birisidir ve amacı metinsel olarak saklanan verilerden (e-postalar, incelemeler, düz metinler, web sayfaları, raporlar, makaleler ve resmi belgeler gibi) [6] anlamlı bilgi çıkarımı yapmaktır [7-8]. Diğer bir ifadeyle, metin madenciliği geniş hacimdeki metin içeriklerinin ana eğilimlerini çıkarmak ve farklı konulardaki uğraşıları analiz etmek için, süreçleri otomatikleştirmeyi mümkün kılan bir tekniktir [9]. Veri madenciliğinde veri seti belli olan, düzgün oluşturulmuş veri tabanları kullanılırken, metin madenciliğinde doğal dil işleme kullanılarak elde edilen düzgün metinler kullanılmaktadır.

Bu çalışmada metin madenciliği teknikleri ile haber makalelerinin önceden tanımlanmış kategorilere göre ayrılması sağlanmıştır. İnternetin gelişmesiyle birlikte insanlar gazetelerden haber okumak yerine dijital haberleri okumayı tercih eder hale gelmişlerdir [10]. Haberciler yazılarını hazırladıktan sonra içeriklerine göre daha önceden

[*] Sorumlu yazar: emrahaydemir@sakarya.edu.tr. Yazarların ORCID Numarası: [1] 0000-0002-8380-7891, [2] 0000-0003-3200-1609, [3] 0000-0002-1425-4664

belirledikleri kategorilere yerleştirirler. Esiyok ve ark. [11] yılında yaptıkları bir çalışmada bu kategorilerin önemini ve haber okuyucularının aynı kategoriye ait haberleri okuma eğiliminde olduklarını göstermiştir. Bu yüzden haberlerin kategorilere ayrılması, okuyucuların hazırlanan haberleri okuması açısından oldukça önemlidir. Geliştirilen model habercilerin hazırladıkları haber makalelerinin kategorilere ayrılmasında bir yardımcı araç olarak kullanılabilecektir.

2. Amaç ve Hedef

Bu çalışmada metin madenciliği teknikleri kullanılarak metinsel veri yığını olarak alınmış haber içeriklerinin önceden tanımlanmış haber kategorilerine göre ayrılması amaçlanmıştır. Yaşam, dünya, ekonomi, kültür-sanat, magazin, otomobil, spor, teknoloji olmak üzere toplam sekiz kategori üzerinde çalışılmıştır. Geliştirilen modelin haberciler tarafından kullanılabilmesi için sınıflandırma başarısı olarak en az %90 başarı oranının elde edilmesi hedeflenmiştir. Metinsel ifadeleri kullanarak sınıflandırma işlemi; doğal dil işleme (DDİ) gibi dil alanının en temel problemlerinden biri olarak sayılabilir [12]. Türkçe dilinin kendine özgün yapısından dolayı metinsel verilerle çalışmak ve anlamlı sonuçlar elde etmek ise oldukça zordur [13]. Bundan dolayı bu çalışmanın sonraki yapılacak DDİ çalışmaları için faydalı olması hedeflenmektedir.

3. Benzer Çalışmalar

Usmani ve Shamsi [14] yapmış oldukları bir çalışmada haber başlıkları sınıflandırma algoritması geliştirilmişlerdir. Bu algoritmada basit DDİ teknikleri ile %88 başarıya ulaşılmışlardır. Acı ve Çırak, [15] 2019 yılında Türkçe haber metinlerini Konvolüsyonel Sinir Ağları (KSA) ve Kelime Vektörü (Word2Vec) algoritmasını kullanılarak %93,3 oranda başarıyla sınıflandırmıştır. 2018 yılında yapılan bir çalışmada yazılan haberin yapısal özelliklerine göre, haber makaleleri kategorilere ayrılmıştır. Sınıflandırma aracı olarak Destek Vektör Makineleri (DVM) kullanılmıştır. Ters piramit yapısına göre hazırlanan haberleri tespit etmede %81,7 başarı sağlarken, Martini yapısına göre haberleri tespit etmede %19,1 oranında başarıya ulaşmışlardır [16]. 2017 yılında yapılan bir çalışmada haber başlıkları kullanılarak ilgili haberin kategorisi belirlenmiştir. Farklı tekniklerle ortalama %76 - %79 arasında başarı elde etmişlerdir [17]. 2017 yılında yapılan bir diğer çalışmada ise farklı haber sitelerinden 4 farklı kategoriye ait 20'şer haber alınarak toplam 80 haber metni toplanmıştır. Toplanan haberlerin 60 tanesi eğitim, 20 tanesi de test için kullanılmıştır. Topladıkları tüm haberleri başarılı bir şekilde sınıflandırmışlardır [18]. Toraman ve ark. [19] 2011 yılında yaptıkları çalışmada; C4.5, en yakın komşu (eYK), Naive Bayes (NB) and DVM (Radyal temel işlevli çekirdek ve Poly) algoritmaları kullanılarak iki farklı veri seti üzerinde çalışmışlar ve sırasıyla %83,3 ve %87,5 başarı elde etmişlerdir. 2010 yılında yapılan farklı bir çalışmada ise haber içerikleri kullanılarak sınıflandırma yapılmış ancak sadece belirli kategorilerde başarılı sonuçlar elde edilmiştir [20]. 2007 yılında yapılan bir çalışmada haberleri okuyan kişilerin duygusal durumlarına göre haberleri daha önceden belirlenen 8 sınıfa ayırmışlardır. Veri olarak haber başlıkları, haber saati, haber kategorisi, haberin meydana geldiği yer ve haber türü alınmıştır. Sınıflandırma aracı olarak ise DVM kullanılmıştır. Okuyucunun duygusal durumuna göre haberleri ayırmada %71,26 başarılı olmuşlardır [21]. 2005 yılında Maksimum Marj Etiketleme (MME) tabanlı yeni bir algoritma ile haber konusuna göre kategorilere ayırma işlemi yapılmıştır. 5000 adet rastgele seçilmiş haber makalelerine ait konular üzerinde yaptıkları çalışmada, haber sayısı arttıkça başarı oranı oldukça düşmüştür [22]. 1999 yılında yapılan bir çalışmada ise yazılan haberlerin anahtar kelimeleri kullanılarak sınıflandırma yapılmış ve sadece belirli kategorilerde %70 - %76,7 arasında başarı elde etmişlerdir [23].

Haber kategorisinin tespiti ile ilgili literatürde yer edinmiş çalışmalar ve okuyucuların aynı kategoriye ait haberleri okuma eğilimleri [11], haber kategorisi seçiminin önemini göstermektedir. Buraya konu olan benzer çalışmaların, bu çalışma ile arasındaki farklılıklar sonuç bölümünde incelenmiştir.

4. Yöntem

Bu çalışmada Türkçe dilinde çevrimiçi (online) olarak haber yapan bir internet sitesinden sekiz kategoriye ait 2248 adet haber metinleri toplanmıştır. C# programlama dili kullanılarak haber metinleri otomatik olarak elde edilmiştir. Haber metinlerinden tüm görseller, bağlantılar, reklam ifadeleri, birden fazla boşluğun yan yana olması gibi gereksiz metinler çıkarılmıştır ve ham metin kayıt altına alınmıştır. Ayrıca metinsel ifadede geçen tüm hipermetin işaretleme dili (Hypertext Markup Language) etiketleri ve içerikleri temizlenmiştir. Bunun haricinde veri setine herhangi bir ön işlem uygulanmamış; bağlaçlar, soru ekleri gibi tek başına anlamsız olan kelimeler çıkartılmamıştır.

Toplanan veriler Weka programı kullanılarak analiz edilmiştir. Weka programı arff uzantılı dosyalar üzerinde işlem yapması nedeniyle veriler haber metni ve kategorisi olacak şekilde iki sütunlu ayarlanmış arff uzantılı olarak düzenlenmiştir. Veriler başka araştırmacılar tarafından test amaçlı kullanılabilmesi için açık kaynak olarak herkesin kullanımına açılmıştır [24]. Metinlerin analizinde ve sınıflandırılmasında, varsayılan parametreler ile sınıflandırma analizi yapan ve yaygın olarak kullanılan algoritmalardan biri olan Çok Terimli Naive Bayes Algoritması (ÇTNBA) [25] ile rastgele orman (RO) (Random Forest) [26] algoritmaları kullanılmıştır. Çalışma içerisinde kullanılan bu iki yöntemin sonuçları verilmiş ve karşılaştırmaları yapılmıştır.

Kappa istatistiği iki değer arasındaki karşılaştırmalı uyuşmanın güvenliğini ölçen bir istatistik yöntemdir [32]. Hesaplanırken iki farklı olasılık hesaplanır. Bunlar iki değerleyici için gözlemlenen uyuşmaların toplama orantısı $Pr(a)$ ve bu uyuşmanın şans eseri ortaya çıkma olasılığı $Pr(e)$'dir. $Pr(a)$ iki değerlendirici için gözlemlenen uyumların toplam orantısı iken, $Pr(e)$ bu uyumun şansa bağlı ortaya çıkma olasılığıdır. Bu iki olasılık üzerinden kappa istatistiği için kullanılan formül Denklem 1'de bulunmaktadır.

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

Hata matrisinin analizlerde nasıl kullanılacağını anlamak için öncelikle aşağıdaki tanımların bilinmesi gereklidir.

- Doğru Pozitif (DP) (True Positive): Gerçek değeri pozitif olup pozitif olarak tahmin edilenler.
- Yanlış Negatif (YN) (False Negative): Gerçek değeri pozitif olup negatif olarak tahmin edilenler.
- Yanlış Pozitif (YP) (False Positive): Pozitif olarak tahmin edilmiş gerçek değeri negatif olanlar.
- Doğru Negatif (DN) (True Negative): Negatif olarak tahmin edilmiş ve gerçek değeri negatif olanlar.

Hata matrisi yukarıda belirtilen tanımlar doğrultusunda oluşturulur. Aşağıdaki Tablo 1'de pozitif sınıf kabulü için örnek bir hata matrisi verilmiştir.

Tablo 1. Örnek Hata Matrisi Tablosu

		Tahmin Değerleri		
		a	b	
Gerçek Değerler	a	DP	YN	Gerçek a toplamı
	b	YP	DN	Gerçek b toplamı
		Toplam Pozitif	Toplam Negatif	

Kategorik değerlerin başarı analizinde kullanılan bir başka hassasiyet (precision), geri çağırma (recall) ve F-Ölçüsü değerleri Denklem 2,3 ve 4'te bulunan formüller kullanılarak hesaplanmıştır.

$$Hassasiyet\ (p) = \frac{dp}{dp+yn} \tag{2}$$

$$Geri\ Çağırma\ (r) = \frac{dp}{dp+yp} \tag{3}$$

$$F - Ölçüsü = 2 \times \frac{Kesinlik \times Hassasiyet}{Kesinlik + Hassasiyet} = 2\frac{pr}{p+r} \tag{4}$$

Denklem 5'te bulunan doğruluk oranı ise doğru tahmin edilen sınıfların toplam örneklem sayısına oranıdır. Başarı ölçütü olarak en çok kullanılan formül olmakla birlikte çok kolay olarak hesaplanır.

$$Doğruluk\ Oranı = \frac{dp+dn}{dp+dn+yp+yn} \tag{5}$$

NB sınıflama algoritması; Bayes teorisi tabanlı [27] anlaşılması kolay ve benzerlerine kıyasla oldukça hızlı çalışan bir algoritmadır. NB, makine öğrenmesi, veri madenciliği ve metin madenciliği için kullanılan en etkili tümevarımsal öğrenme algoritmalarından biri olmasının yanı sıra hem tahmin edici hem de tanımlayıcı bir sınıflama tekniğidir [28].

RO; Leo Breiman ve Adele Cutler tarafından geliştirilen [29] ve karar ağaçları için tanımlanmış olmasına rağmen, tüm sınıflandırıcılar için kullanılabilen bir algoritmadır [30]. RO içerisinde oylama metodunu bulunur ve

birçok karar ağacının biraya gelmesiyle oluşur ve bireysel ağaçlar tarafından oylanarak kazanan sınıf belirlenir. Rastgele orman yönteminin önemli bir avantajı, çok sayıda girdi değişkenlerini ele almasıdır [30].

Rastgele orman algoritması kullanılmadan önce metinden dizi kelime (MDK) (String To Word Vector) filtresi kullanılarak haber metninden elde edilen haber metinleri, kelime dizisi şeklinde sütunlara dönüştürülmüştür. Bu sayede farklı öznitelikler elde edilmesinin önü açılan veri yapısına, tüm sınıflandırma algoritmalarının uygulanması mümkün hale gelmiştir. Çalışmaya konu olan her iki sınıflandırma algoritması kullanılarak sınıflandırma işlemi yapılmış ve elde edilen veriler Bulgular kısmında paylaşılmıştır. Toplanan haber metni sayısına göre çıkan sonuçların güvenilirliği için çapraz doğrulama tekniği olan 10-katlı çapraz doğrulama (cross validation) tekniği kullanılmıştır. Ayrıca elde edilen sonuçlar iki veya daha fazla gözlemci arasındaki uyumun güvenilirliğini ölçen bir istatistik yöntemi olan ve literatürde en çok tercih edilen [31] kappa katsayısı hesaplanarak kontrol edilmiştir.

5. Veri Seti ve Bulgular

Birbirine yakın sayıda haber içeren 2248 haber metni toplanmıştır. Tablo 2'de her bir haber kategorisinden toplanan haberler ve bu haberlerin ortalama sözcük sayısı görülmektedir. Her bir haber metni ortalama 1586 sözcükten oluşmaktadır. En uzun haberler Kültür-Sanat kategorisine ait iken en kısa haberler ise Magazin kategorisinde olmuştur. Bu duruma sebep olarak magazin haberlerinin ağırlıklı olarak fotoğraflardan oluşması gösterilebilir.

Tablo 2. Kategorilerine Göre Haber Sayıları

Kategoriler	Toplam Haber Sayısı	Ortalama Sözcük Sayısı
Yaşam	292	1018
Dünya	249	1450
Ekonomi	261	2127
Kültür-Sanat	183	3140
Magazin	315	606
Otomobil	303	2056
Spor	315	1409
Teknoloji	330	1573
Genel Toplam/Ortalama	**2248**	**1586**

ÇTNBA kullanılarak veriler analiz edildiğinde %95,24 oranında başarılı sınıflandırma yapılmıştır. 2248 haberden 2141 tanesi doğru sınıfta tahmin edilmiş ve 0,94 kappa istatistiği elde edilmiştir. Kappa değerinin 1'e yakın olması, tahmin ve gerçek değer arasında uyuşmanın yüksek olduğunu göstermektedir. ÇTNBA için oluşan karışıklık matrisi Tablo 3'te verilmiştir.

Tablo 3. ÇTNBA için karışıklık matrisi

Kategoriler	Yaşam	Dünya	Ekonomi	Kültür-Sanat	Magazin	Otomobil	Spor	Teknoloji
Yaşam	281	0	0	0	11	0	0	0
Dünya	4	240	0	0	5	0	0	0
Ekonomi	7	2	248	0	4	0	0	0
Kültür-Sanat	10	0	0	130	43	0	0	0
Magazin	2	0	0	0	313	0	0	0
Otomobil	0	0	0	0	5	298	0	0
Spor	2	2	0	0	2	0	309	0
Teknoloji	0	0	0	0	8	0	0	322

ÇTNBA ile yapılan analiz sonuçlarına yönelik Doğru Pozitif Oranı (DP), Yanlış Pozitif Oranı (YP) ve F-Ölçüsü gibi değerler Tablo 4'te verilmiştir.

Tablo 4. ÇTNBA için Analiz Sonuçları

	DP Oranı	YP Oranı	Hassasiyet	Geri Çağırma	F-Ölçüsü	MCC	ROC Alanı	PRC Alanı	Sınıf
	0,962	0,013	0,918	0,962	0,940	0,931	0,999	0,991	Yaşam
	0,964	0,002	0,984	0,964	0,974	0,970	0,999	0,996	Dünya
	0,950	0,000	1,000	0,950	0,974	0,972	0,999	0,995	Ekonomi
	0,710	0,000	1,000	0,710	0,831	0,832	0,990	0,946	Kültür-Sanat
	0,994	0,040	0,801	0,994	0,887	0,873	0,996	0,957	Magazin
	0,983	0,000	1,000	0,983	0,992	0,990	1,000	1,000	Otomobil
	0,981	0,000	1,000	0,981	0,990	0,989	0,999	0,996	Spor
	0,976	0,000	1,000	0,976	0,988	0,986	0,999	0,998	Teknoloji
Ağırlıklı Ortalama	0,952	0,008	0,960	0,952	0,952	0,948	0,998	0,987	

Elde edilmiş bu başarılı sınıflandırma oranı biraz daha geliştirilmek istenmiş ve verilere MDK filtresi uygulandıktan sonra elde edilen 5774 sütunlu veriler rastgele orman algoritması ile analiz edilmiştir. Bu analizde %99,86 oranında başarılı sınıflandırma yapılmış ve 2248 veriden yalnızca üç tanesi yanlış sınıflandırılmış ve 0,99 kappa istatistiği elde edilmiştir. Rastgele orman algoritması için oluşan karışıklık matrisi aşağıdaki Tablo 5'te verilmiştir.

Tablo 5. Rastgele orman için karışıklık matrisi

Kategoriler	Yaşam	Dünya	Ekonomi	Kültür-Sanat	Magazin	Otomobil	Spor	Teknoloji
Yaşam	292	0	0	0	0	0	0	0
Dünya	0	249	0	0	0	0	0	0
Ekonomi	0	0	261	0	0	0	0	0
Kültür-Sanat	0	0	0	183	0	0	0	0
Magazin	0	0	0	0	315	0	0	0
Otomobil	0	0	0	0	0	303	0	0
Spor	0	0	0	0	0	0	315	0
Teknoloji	0	0	0	0	3	0	0	327

Rastgele Orman ile yapılan analiz sonuçlarına yönelik Doğru Pozitif Oranı (DP), Yanlış Pozitif Oranı (YP) ve F-Ölçüsü gibi değerler Tablo 6'da verilmiştir.

Tablo 6. Rastgele Orman için Analiz Sonuçları

	DP Oranı	YP Oranı	Hassasiyet	Geri Çağırma	F-Ölçüsü	MCC	ROC Alanı	PRC Alanı	Sınıf
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	Yaşam
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	Dünya
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	Ekonomi
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	Kültür-Sanat
	1,000	0,002	0,991	1,000	0,995	0,994	1,000	1,000	Magazin
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	Otomobil
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	Spor
	0,991	0,000	1,000	0,991	0,995	0,995	1,000	1,000	Teknoloji
Ağırlıklı Ortalama	0,999	0,000	0,999	0,999	0,999	0,998	1,000	1,000	

Karışıklık matrisinde köşe değerlerin kategori sayısına yakınlığı ilgili kategorinin yüksek tahmin başarısına sahip olduğunu göstermektedir. Tablo 3 ve 5'te bulunan karışıklık matrisleri incelendiğinde incelendiğin de bazı

köşe değerlerinin olabilecek en yüksek değer olduğu diğerlerinin ise oldukça başarılı olduğu görülmektedir. En çok hatalı sınıflandırma işleminin Magazin ve Yaşam alanlarında olduğu görülmektedir. Bunun temel sebebi olarak alanların birbirlerine oldukça yakın olmasından dolayı fazla benzer kelimeler içermesi gösterilebilir. Spor, Otomobil, Ekonomi, Teknoloji ve Kültür-Sanat gibi birbirine uzak alanların sınıflandırılmasında ise hiçbir hata elde edilmemiştir. Bunun temel sebebi olarak bu alanlarda kullanılan kelimelerin alana özgü olmasından dolayı farklılık göstermesidir. Ayrıca elde edilen Kappa değerlerinin 1'e oldukça yakın olması [32] değerlendirmedeki uyumun oldukça yüksek olduğunu göstermektedir.

5. Tartışma ve Sonuç

Veri seti olarak kullanılan haber metinlerinin sayısının artırılması, farklı dillerdeki metinlerin kullanılması ve daha fazla kategoriye göre incelenmesiyle elde edilecek sınıflandırma başarılarının karşılaştırılması faydalı olacaktır. Türkçe dilinin özgün yapısından dolayı veri seti elde edilirken mümkün olduğu kadar farklı yazarların haberlerinin toplanması sınıflandırma başarısının doğruluğunu artıracaktır.

Bu çalışmanın, benzer çalışmalar ile karşılaştırması Tablo 7'de verilmiştir. Tablo incelendiği zaman bu çalışmaya konu olan yöntemler ile elde edilen sonuçların oldukça başarılı olduğu ve metinsel içeriği hazırlanan bir haberin önceden belirlenmiş kategorilere göre sınıflandırılmasında kullanılabileceği görülmektedir.

Tablo 7. Benzer çalışmalar ile karşılaştırılması

Benzer Çalışmalar	**Kullanılan veri sayısı**	**Başarı Oranı**	**Bu çalışma ile farklılıkları**
Usmani ve Shamsi [14]	2500000	%88	Haber başlıkları kullanılarak sınıflandırma yapılmıştır.
Acı ve Çırak [15]	600	%93,3	KSA ve Kelime Vektörü Kullanılarak sınıflandırılma yapmışlardır.
Dai ve ark. [16]	853	%81,7	Metin Türkçe değildir ve Yazının yapısal özelliğini veri olarak kullanmıştır.
Qiu ve ark.[17]	12000	%76 – %79	Metin Türkçe değildir ve Haber başlıklarını veri olarak kullanmıştır. Farklı kategorilerde farklı sonuçlar elde etmiştir.
Başkaya ve Aydın [18]	80	%100	İncelenen veri sayısı oldukça az olmasının yanı sıra, sadece 4 kategori üzerinde sınıflandırma yapılmıştır.
Toraman ve ark. [19]	7540	%83,3 – %87,5	C4.5, en yakın komşu (eYK), Naive Bayes (NB) and DVM algoritmaları kullanılarak farklı sonuçlar elde etmişler ve bu sonuçları karşılaştırmışlardır. Toplam 6 farklı kategori için çalışma yapılmıştır ancak kategoriler birbirinden oldukça uzak seçilmiştir.
Krishnalal ve ark. [20]	400	%80,22 - %96,34	Sadece 4 kategoriye bölme işlemi yapılmıştır.
Lin ve ark. [21]	1200	%80,22 – %96,34	Metin Türkçe değildir. Sadece birkaç kategoride yüksek sonuçlar elde edilmiştir.
Kazawa ve ark. [22]	5000	Veri sayısına göre değişmektedir.	Haber konularından, haber kategorisi tespit etmeye yönelik olup, veri sayısı artıkça başarı oranı düşmektedir.
Jo [23]	1000	%70 – %76,7	Metin Türkçe değildir ve yazılan haberlerin anahtar kelimeleri veri olarak kullanmıştır. Farklı kategorilerde farklı sonuçlar elde etmiştir.
Bu çalışma	**2248**	**%99,86**	

Ortalama 1586 sözcükten oluşan 2248 haber metni (Türkçe dili) veri olarak kullanılmıştır. Toplanan veriler ÇTNBA kullanılarak analiz edildiğinde %95,24 oranında başarılı sınıflandırma ve 0,94 kappa istatistiği elde

edilmiştir. Sonrasında bu verilere MDK filtresi uygulanıp RO algoritması ile analiz edilmesi sonucunda %99,86 oranında başarılı sınıflandırma ve 0,99 kappa istatistiği elde edilmiştir.

Kaynaklar

[1] Doğan K, Arslantekin S, Büyük veri: önemi, yapısı ve günümüzdeki durum. Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi 2016; 56(1): 15-36.

[2] Gautam P, Singh YP, Shaikh P, Significance and Importance of Data Mining for Marketing Analysis in Finance. Banking Sectors, Int. J. Appl. Res. Sci. Eng 2017; 26–29.

[3] Khedr AE, Salama SE, Yaseen N, Predicting stock market behavior using data mining technique and news sentiment analysis. Int. J. Intell. Syst. Appl. 2017; 9(7): 22-30.

[4] Martinez-Martin N, Insel TR, Dagum P, Greely HT, Cho MK, Data mining for health: staking out the ethical territory of digital phenotyping. npj Digit. Med. 2018; 1(1): 1-5.

[5] Bustince H, Herrera F, Montero J. Fuzzy Sets and Their Extensions: Representation, Aggregation and Models. 1th ed. Springer-Verlag Berlin Heidelberg, 2008.

[6] Bach MP, Krstič Ž, Seljan S, Turulja L. Text mining for big data analysis in financial sector: A literature review. Sustain 2019; 11(5): 2019.

[7] Alsaidi SA, Sadiq AT, Abdullah HS. English poems categorization using text mining and rough set theory. Bull. Electr. Eng. Informatics 2020; 9(4): 1701-1710.

[8] Doğan K, Arslantekin S. Elektronik Belge Yönetimi, Dijital Arşivleme Sistemleri ve Büyük Veri. Bilgi Sistemleri ve Bilişim Yönetimi: Beklentiler ve Yeni Yaklaşımlar, Ankara Üniversitesi Basımevi, 2017; 65-80.

[9] Monino JL, Sedkaoui S. Big Data, Open Data and Data Development. 3nd ed. London: ISTE Ltd., 2016.

[10] Liu C, Wang W, Zhang Y, Dong Y, He F, Wu C. Predicting the Popularity of Online News Based on Multivariate Analysis. IEEE International Conference on Computer and Information Technology (CIT); 21-23 Agust 2017; Helsinki, Finland.

[11] Esiyok C, Kille B, Jain BJ, Hopfgartner F, Albayrak S. Users' reading habits in online news portals. 5th Information Interaction in Context Symposium; 26-29 August 2014; New York, U.S.A.

[12] Sukiennik N, Hui P. Inflo: News Categorization and Keyphrase Extraction for Implementation in an Aggregation System. ArXiv; 2018; abs (1812.03781).

[13] Yüksel A, Tan G. Metin Madenciliği Teknikleri ile Sosyal Ağlarda Bilgi Keşfi. Mühendislik Bilimleri ve Tasarım Dergisi 2018; 6(2): 324-33.

[14] Usmani S, Shamsi JA. News Headlines Categorization Scheme for Unlabelled Data. In 2020 International Conference on Emerging Trends in Smart Technologies (ICETST); 26 – 27 March 2020; Karachi, Pakistan: IEEE. pp. 1-6

[15] Acı Çİ, Çırak A, Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması. International Journal of InformaticsTechnologies 2019; 12(3).

[16] Dai Z, Taneja H, Huang R. Fine-grained structure-based news genre categorization. 2018 Events and Stories in the News Workshop; 20-21 August 2018; New Mexico, U.S.A.

[17] Qiu X, Gong J, Huang X. Overview of the NLPCC 2017 shared task: Chinese news headline categorization. National CCF Conference on Natural Language Processing and Chinese Computing; 8-12 November 2017; Dalian, China.

[18] Başkaya F, Aydin İ. Haber metinlerinin farklı metin madenciliği yöntemleriyle sınıflandırılması. International Artificial Intelligence and Data Processing Symposium (IDAP); 1-5, September 2017; Malatya, Turkey.

[19] Toraman C, Can F, Koçberber S. Developing a text categorization template for Turkish news portals. International Symposium on Innovations in Intelligent Systems and Applications, June 2011; 379-383.

[20] Krishnalal G, Rengarajan SB, Srinivasagan KG, A new text mining approach based on HMM-SVM for web news classification. International Journal of Computer Applications, 2010; 1(19): 98-104.

[21] Lin KHY, Yang C, Chen HH, What emotions do news articles trigger in their readers?. 30th annual international ACM SIGIR conference on Research and development in information retrieval; 23-27 July 2007; Amsterdam, Holland.

[22] Kazawa H, Izumitani T, Taira H, Maeda E. Maximal margin labeling for multi-topic text categorization. Advances in neural information processing systems 2005; 649-656.

[23] Jo TC. "Text categorization with the concept of fuzzy set of informative keywords". 1999 IEEE International Fuzzy Systems Conference Proceedings; 22-25 August 1999; 99CH36315(2): 609-614.

[24] https://websiteyonetimi.ahievran.edu.tr/_Dosyalar/Genel/HaberMetinleri.rar, E.T.:01.03.2021.

[25] https://weka.sourceforge.io/doc.dev/weka/classifiers/bayes/NaiveBayesMultinomialText.html, E.T.:01.03.2021.

[26] https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/RandomForest.html, E.T.:01.03.2021.

[27] Arpacı SA, Kalıpsız O. Yazılım Hata Sınıflandırmasında Farklı Naive Bayes Tekniklerin Kıyaslanması. Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi 2018; 7(1): 1-13.

[28] Aydoğan E. Veri Madenciliğinde Sınıflandırma Problemleri İçin Evrimsel Algoritma Tabanlı Yeni Bir Yaklasım: Rough-Mep Algoritması. Doktora tezi, Gazi Üniversitesi, 2008, Ankara.

[29] Skurichina M, Duin RPW. Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis and Applications 2002; 5(2): pp. 121–135.

[30] Korkem E, Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest ve Naıve Bayes Sınıflama Yöntemleri Yaklaşımı. Yüksek Lisans tezi, Hacettepe Üniversitesi, 2013, Ankara.
[31] Zec S, Soriani N, Comoretto R, Baldi I. Suppl-1, M5: high agreement and high prevalence: the paradox of Cohen's Kappa. The open nursing journal 2017; 11(1).
[32] Cohen JA. Coefficient of Agreement for Nominal Scales", Educational and Psychological Measurement 1960; 20(1): 37-46.