SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
VOL. 2, NO. 1, APRIL 2019

Makine Öğrenmesi ile Ürün Kategorisi Sınıflandırma

Serap Kazan[1], Hakan Karakoca[2]
[1]SAÜ, Bilgisayar Mühendisliği; scakar@sakarya.edu.tr; *+90 264 295 59 04*
[2]SAÜ, Bilgisayar Mühendisliği; hakan.karakoca@ogr.sakarya.edu.tr

Received: 06/02/2019; Revision; 11/02/2019 Accepted; 20/03/2019 Published online; 25/04/2019

Öz

Teknolojinin ilerlemesi ve internetin gelişmesi ile beraber günümüzde bilginin gücü de ön plana çıkmıştır. Bununla beraber internet dünyasında bilgi kirliliği ve karmaşası ortaya çıkmaya başlamıştır. Bu karmaşadan anlamlı verilerin çıkartılması ve yorumlanabilmesi için makine öğrenmesi algoritmalarından yararlanılabilir. Bu çalışmada yazı formunda girilen açıklamanın kategori bilgisine ulaşılması amaçlanmıştır. Bir e-ticaret sitesinden ürün bilgileri etiketlenerek veri seti elde edilmiştir. Toplanan bu veri seti makine öğrenmesi algoritmalarıyla model eğitimi gerçekleştirilmiş ve 9 farklı katagoriye ayırmak için doğru tahminleme yapması amaçlanmıştır. Bu eğitim sırasında Random Forest, Karar Ağacı, Multinominal Naive Bayes (Multinominal NB), Lojistik Regresyon, Destek Vektör Makineleri (DVM) ve Yapay Sinir Ağları (YSA) sınıflandırıcıları kullanılmış ve çıkan sonuçlar hata matrisleri gösterilerek tablolarla karşılaştırılmıştır.

Anahtar Kelimeler: makine öğrenmesi, ürün kategorisi sınıflandırma

Product Category Classification with Machine Learning

Abstract

With the advancement of technology and the development of the internet, the power of knowledge has come to the fore. However, in the internet world, information pollution and chaos started to emerge. Machine learning algorithms can be used to extract and interpret meaningful data from this complex. In this study, it is aimed to reach the category information of the explanation entered in the form of text. Product information from an e-commerce site was obtained by labeling the data set. This data set is modeled by machine learning algorithms and it is aimed to make accurate estimation to divide into 9 different categories. During this training, Random Forest, Decision Tree, Multinominal Naive Bayes (Multinominal NB), Logistic Regression, Support Vector Machines (SVM) and Artificial Neural Networks (ANN) classifiers were used and the results were compared with the tables by showing the confussion matrix.

Keywords: machine learning, product category classification

1. Giriş

İnternet siteleri, kullanıcılar, çevrimiçi depolama sağlayıcıları ve sosyal medyadaki hızlı artış günlük olarak mevcut veri miktarını arttırmaktadır. Bu veriler genellikle yapılandırılmamaktadır ve bu verilerin yönetilmesi ve düzenlenmesi için doğru bir otomatik metin kategorizasyon sistemi gereklidir. Bu nedenle, metin kategorizasyonu önemli bir araştırma konusu olmaya devam etmekte ve araştırma topluluğu ve endüstrisinden büyük ilgi görmektedir.

Aliwy ve arkadaşları, dönüşüm temelli yaklaşımlar kullanarak Arapça çoklu etiketli metin kategorizasyonu üzerine bir çalışma yürütmüşlerdir. Değerlendirme için beş kategoriye (“sanat”, “spor”, “politika”, “ekonomi” ve “bilim”) dağıtılmış 10.000 belgeden oluşan bir veri seti kullanılmıştır. Veri seti BBC Arabic haber portalından toplanmış ve her haber makalesinin etiketleri kategori olarak kullanılmıştır. Sınıflandırma için DVM, *k*-En yakın komşu (*k*-NN) ve Random Forest gibi algoritmalar kullanılmıştır [1].

Alshalabi ve arkadaşları, Malayca metinlerin makine öğrenmesi algoritmaları ile otomatik olarak kategorize edilmesi üzerine bir çalışma yapmışlardır. Çalışmada iki özellik seçme yöntemi (Bilgi kazancı (IG) ve Ki-kare) ve üç makine öğrenmesi yöntemi (*k*-NN, Naive Bayes (NB) ve N-gram) kullanılmıştır [2].

Hmeidi ve arkadaşları, Arapça metinlerin sınıflandırılması ile ilgili bir çalışma yapmışlardır. Çalışmada NB, DVM, *k*-NN ve Karar Ağaçları gibi sınıflandırıcılar kullanılarak metinler ekonomik, politik ve spor gibi katagorilere ayırılmıştır [3].

Aggarwal ve Zhai, metin sınıflandırma algoritmaları ile ilgili bir araştırma çalışması yapmışlardır. Çalışmada Karar Ağaçları, Kural tabanlı sınıflandırıcılar, DVM, YSA ve Bayes gibi sınıflandırıcılarla ilgili ayrıntılı araştırma yapmışlardır [4].

Bu çalışmada bir e-ticaret sitesinden alınan veriler, Random Forest, Karar Ağacı, Multinominal NB, Lojistic Regresyon, DVM ve YSA sınıflandırıcıları kullanılarak 9 farklı katagoriye sınıflandırılmıştır. Sınıflandırıcıların başarım oranları tablolarla karşılaştırılmıştır.

2. Kullanılan Yöntemler

2.1 Naive Bayes

Naive Bayes (NB) sınıflandırma Bayes teoremine dayanır ve hedef sınıfı için verilen değerin gerçekleşme olasılığının ne olduğunu bildirir (Denklem 1) [5].

$$\boldsymbol{P}(\mathrm{c}|\boldsymbol{x}) = \frac{\boldsymbol{P}(\mathrm{c}|\boldsymbol{x})\,\boldsymbol{P}(\boldsymbol{c})}{\boldsymbol{P}(\boldsymbol{x})} \tag{1}$$

Denklem (1)'de; *c,* tahmin edilmeye çalışılan sınıf, *x,* tahmin eden sınıf, *P(c|x), x* olayı gerçekleştiğinde *c* olayının gerçekleşme olasılığı, *P(x|c), c* olayı gerçekleştiğinde *x* olayının gerçekleşme olasılığı, *P(c), c* olayının gerçekleşme olasılığı, *P(x), x* olayının gerçekleşme olasılığıdır.

2.1.1. Multinominal NB

Multinominal NB yöntemi çok terimli Naive Bayes sınıflandırıcısını oluşturmak için kullanılabilir. Hızlı, kolay uygulanabilir ve oldukça verimlidir. Multinomial Naive Bayes belgedeki kelimelerin dağılımını çok terimli olarak modeller. Belge kelime dizisi olarak ele alınır ve her kelimenin pozisyonunun diğerinden bağımsız olarak oluşturulduğu varsayılır. Naive Bayes'in çok terimli versiyonu olan Multinominal NB, Rennie ve Shih tarafından ortaya atılmış, analiz edilmiş ve geliştirilmiştir [6].

2.2. Lojistik Regresyon

Lojistik regresyon, ikili sınıfları öngörmek için istatistiksel bir yöntemdir. Lojistik regresyon, yalnızca iki değere sahip olabilen bir sonucun olasılığını öngörür. Tahmin, bir veya birkaç öngörücünün (sayısal ve kategorik) kullanımına dayanır. Doğrusal regresyon evet/hayır, var/yok gibi binary (ikili) sistemde ifade edilebilecek değerler için uygun değildir. Çünkü, 0 ve 1 aralığının dışında değer tahmin edebilir. Lojistik regresyon, 0 ile 1 arasındaki değerlerle sınırlı lojistik eğrisi üretir [7].

2.3. Random Forest

Random Forest sınıflandırıcısı denetimli bir öğrenme algoritmasıdır. Hem sınıflandırma hem de regresyon için kullanılabilir. Aynı zamanda esnek ve kullanımı kolay bir algoritması vardır. Rastgele ormanlar rastgele seçilmiş veri örneklerinde karar ağaçları oluşturur, her ağaçtan tahmin yapar ve oylama ile en iyi çözümü seçer [8].

Dört adımda çalışır;

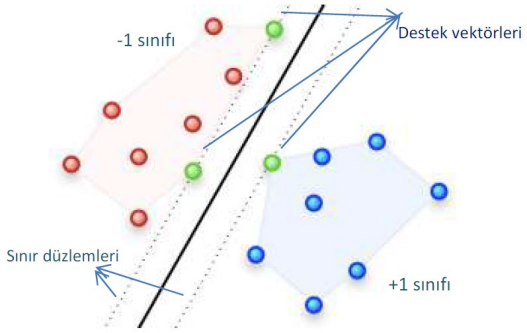
1. Belirli bir veri kümesinden rastgele örnekler seçilir.
2. Her örnek için bir karar ağacı oluşturulur ve her karar ağacından bir tahmin sonucu alınır.
3. Tahmin edilen her sonuç için bir oylama yapılır.
4. Tahmin sonucu, son tahmin olarak en çok oy kullanılarak seçilir.

2.4. Karar Ağacı

Karar ağacı, hem regresyon hem de sınıflandırma problemlerinde kullanılabilen denetimli bir öğrenme algoritması türüdür. Hem kategorik hem de sürekli giriş ve çıkış değişkenleri için çalışır. Karar Ağacı olası tüm eylem seçeneklerini, bu eylem seçeneklerine etkisi olabilecek tüm olası faktörleri ve tüm bu faktörlere dayanan her bir olası sonucu, verilere bağlı olarak değerlendiren, çizgi, kare, daire gibi geometrik semboller kullanımı yoluyla karar vericiye problemi anlamada kolaylık sağlayan grafiksel bir teknik olarak tanımlanabilir. Karar Ağacı, grafik gösterimi ile problemin tüm yönlerini ayrıntılı olarak ortaya koymaktadır. Herhangi bir karar problemi için kullanılabilen Karar Ağacı tekniği özellikle birden fazla kararın ardışık olarak verilmesini gerektiren karar problemlerinin gösteriminde çok kullanışlıdır [9].

2.5. Destek Vektör Makineleri

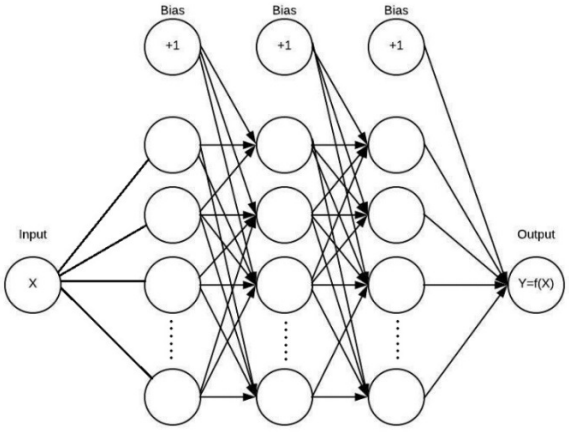
Destek Vektör Makineleri (DVM) öğrenme, sınıflandırma, kümeleme, yoğunluk tahmini ve son olarak da veriden regresyon kuralları üretmek için kullanılan eğitme algoritmasıdır. DVM iki sınıflı ve çok sınıflı sınıflandırma probleminin çözümü için kullanılabilir. DVM veriyi sınıflandırırken sınıfların birbirlerine en yakın örneklerini bularak bu örneklerin (iki sınıfı ayıracak olan) ayırıcı yüzeye dik uzaklıklarını maksimize etmeyi amaçlar. Ayırıcı yüzeyin, veri kümesi üzerindeki başarısı değişmeden birçok farklı alternatifi olabilir. DVM sayesinde ayırıcı yüzey her iki sınıfa da aynı mesafede ve maksimum uzaklıktadır [10]. Şekil 1'de DVM için iki sınıflı problem örneği gösterilmiştir.



Şekil 1 DVM için iki sınıflı problem örneği

2.6. Yapay Sinir Ağları

Yapay Sinir Ağları (YSA)'nın birçok değişik tipi vardır. Bu şalışmada Çok Katmanlı Algılayıcı (Multi-Layer Perceptron-MLP) tipi bir YSA yapısı kullanılmıştır. MLP ağı bir giriş katmanı, 3 gizli katman ve bir çıkış katmanından oluşur. Her katman bünyesinde birçok nöron hücresi bulundurur. Bu çalışmada gizli katmanların her birinde 13 nöron hücresi bulunur. Bu hücreler birbirlerine ağırlıklı bağlantılarla bağlıdır. MLP ağı *n* giriş vektörünü, üzerinde doğrusal olmayan işlemler yaparak bir çıkış vektörüne dönüştürür. Ağın çıkışı bir aktivasyon fonksiyonuna sahip çıkış katmanı tarafından belirlenir. Hesaplanan çıkış değeri ile hedef değer arasındaki fark ortalama karesel hata fonksiyonu olarak tanımlanır. MLP ağının eğitimi, tanımlanan bu hata fonksiyonunun minimizasyonu şeklinde ifade edilen bir süreçtir. Bu süreç içerisinde nöronlararasındaki ağırlıklı bağlantılar optimize edilir. Optimizasyon Eğim Düşümü (Gradient Descent) ve Geri Yayılım (Backpropogation) algoritmaları ile gerçekleştirilir. Hata fonksiyonu karesel ortalama hata şeklinde ifade edildiğinden MLP ağlarının eğitiminde kullanılan eğitim setinin ve test verilerinin dağılımı model tarafından kontrol edilemeyen önemli bir parametredir. Veri seti dağılımının ağ eğitimini etkileyen önemli bir parametere olduğu literatürde gösterilmiştir. MLP tipi sınıflandırıcılar çoklu sınıflandırma problemlerinde kolaylıkla uyarlanabilir [11]. Aktivasyon fonksiyonları olarak bu çalışmada 2 farklı tür denenmiştir (ReLu ve logisticsigmoid). Şekil 2'de veri setinde kullanılan, 3 ara katmana sahip MLP yapısı gösterilmiştir.



Şekil 2 Veri setinde kullanılan, 3 ara katmana sahip MLP yapısı

3. Verilerin Elde Edilmesi ve Sınıflandırılması

Bu çalışmada öncelikle bir e-alışveriş sitesinden veriler elde edilmiştir. Daha sonra veri önişleme aşamasında kullanılan iki farklı özellik çıkartma işlemi olan TF-IDF ve CountVectorizer karşılaştırması yapılmış ve CountVectorizer'ın daha iyi sonuç verdiği gözlemlenmiştir. CountVectorizer bir metin dökümanını terim sayısı matrisine dönüştürür. Tablo 1'de CountVectorizer için kullanılacak olan örnek veri seti, Şekil 4'te CountVectorizer eşsiz kelimeler listesi, Şekil 5'te örnek veri setinin eşsiz kelimeler listesini baz alarak vektör haline getirilmesi gösterilmiştir.

Tablo 1 CountVectorizer için kullanılacak veri seti örnekleri

İndeks	Tip	Büyüklük	Değer
0	str	1	Pierre Cardin 7290 Hamile Lohusa 3 lü Pijama Sabahlık Takımı
1	str	1	Türkçe Hayvan Sesli ve Işıklı Eğitici Oyuncak Mini Çocuk Piyanosu
2	str	1	Fisher Price Yağmur Ormanı Jumperoo
3	str	1	Philips Avent SCF371/60 PP Klasik Yenidoğan Hediye Seti

```
In [33]: print(ct.get_feature_names())
['60', '7290', 'avent', 'cardi', 'erre', 'eğitici', 'fisher', 'hami', 'hayvan', 'hediye',
'işıklı', 'jama', 'jumperoo', 'klasik', 'le', 'lohusa', 'lü', 'mini', 'ormanı', 'oyuncak',
'philips', 'pi', 'piyanosu', 'pp', 'price', 'sabahlik', 'scf371', 'sesli', 'seti', 'takimi',
'türkçe', 've', 'yağmur', 'yenidoğan', 'çocuk']
```

Şekil 4 CountVectorizer eşsiz kelimeler listesi

```
In [34]: print(X.toarray())
[[0 1 0 1 1 0 0 1 0 0 0 1 0 0 1 1 1 0 0 0 0 2 0 0 0 1 0 0 0 1 0 0 0 0 0]
 [0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 1 0 0 1 1 0 0 1]
 [0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0]
 [1 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 0 1 0]]
```

Şekil 5 Örnek veri setinin eşsiz kelimeler listesini baz alarak vektör haline gelmesi

Veri setinde Anne & Bebek, Elektronik, Ev & Yaşam , Giyim & Ayakkabı, Kitap, Film, Müzik, Oyun, Kozmetik & Kişisel Bakım, Mücevher & Saat , Otomotiv & Motosiklet, Spor & Outdoor şeklinde 9 farklı kategorik bilgi, her birinde 250000 adet olacak şekilde eşit olarak tutulmuştur. Toplam 2250000 adet veriden, 1687500 adet veri eğitim için ayrılırken 562500 adet veri test için ayrılmıştır. Tablo 2'de Multinominal NB hata matrisi sonuçları, Tablo 3'de Logistic Regression hata matrisi sonuçları, Tablo 4'de Random Forest hata matrisi sonuçları, Tablo 5'de Karar Ağacı hata matrisi sonuçları, Tablo 6'de DVM hata matrisi sonuçları, Tablo 7'da MLP (aktivasyon='ReLU') hata matrisi sonuçları, Tablo 8'de MLP (aktivasyon='logistic sigmoid') hata matrisi sonuçları verilmiştir. Tablolarda precision (Denklem 2), recall (Denklem 3) ve f1-score (Denklem 4) sonuçları görülmektedir (TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives). Tablo 9'da veri seti üzerinde kullanılan sınıflandırıcıların başarım oranlarının ortalama değerlerinin karşılaştırılması yapılmıştır.

Tablo 2 Multinominal NB sınıflandırıcısı kullanıldığında Hata Matrisi sonuçları

Veri Seti	**Anne&Bebek**	**Elektronik**	**Ev&Yaşam**	**Giyim&Ayakkabı**	**Kitap, Müzik, Film, Oyun**	**Kozmetik& Kişisel Bakım**	**Mücevher&Saat**	**Otomotiv& Motosiklet**	**Spor&Outdoor**
Anne&Bebek	**57302**	405	987	111	1175	2404	94	33	249
Elektronik	245	**61190**	568	118	290	75	76	134	119
Ev&Yaşam	773	448	**58232**	547	781	1104	141	92	177
Giyim&Ayakkabı	214	118	585	**59454**	123	98	210	69	1654
Kitap, Müzik, Film, Oyun	1082	976	1216	105	**57711**	301	135	348	360
Kozmetik&Kişisel Bakım	767	83	779	56	45	**60629**	79	27	41
Mücevher&Saat	324	143	549	116	414	200	**60343**	116	105
Otomotiv&Motosiklet	64	276	264	68	155	122	57	**60864**	638
Spor&Outdoor	482	551	682	3886	412	177	258	1314	**54776**

$$precision = TP/(TP + FP) \quad (2)$$

$$recall = TP/(TP + FN) \quad (3)$$

$$f1 - score = 2 * (recall * precision)/(recall + precision) \quad (4)$$

Tablo 3 Lojistik Regression sınıflandırıcısı kullanıldığında Hata Matrisi sonuçları

Veri Seti	Anne&Bebek	Elektronik	Ev&Yaşam	Giyim&Ayakkabı	Kitap, Müzik, Film, Oyun	Kozmetik& Kişisel Bakım	Mücevher&Saat	Otomotiv& Motosiklet	Spor&Outdoor
Anne&Bebek	**58998**	234	624	76	956	1209	70	29	193
Elektronik	130	**61479**	256	51	343	29	51	87	102
Ev&Yaşam	442	244	**59838**	300	730	453	140	59	184
Giyim&Ayakkabı	97	70	357	**60054**	119	29	55	10	1677
Kitap, Müzik, Film, Oyun	663	337	591	27	**60421**	78	82	100	234
Kozmetik&Kişisel Bakım	602	36	584	24	174	**61187**	48	16	50
Mücevher&Saat	81	45	179	51	206	84	**61874**	16	50
Otomotiv&Motosiklet	27	87	91	36	193	42	34	**61157**	475
Spor&Outdoor	216	166	285	1852	506	44	245	493	**58932**

Tablo 4 Random Forest sınıflandırıcısı kullanıldığında Hata Matrisi sonuçları

Veri Seti	Anne&Bebek	Elektronik	Ev&Yaşam	Giyim&Ayakkabı	Kitap, Müzik, Film, Oyun	Kozmetik& Kişisel Bakım	Mücevher&Saat	Otomotiv& Motosiklet	Spor&Outdoor
Anne&Bebek	**60235**	148	412	60	723	784	57	42	187
Elektronik	132	**61665**	220	66	315	46	27	62	125
Ev&Yaşam	548	227	**59083**	314	903	532	122	62	125
Giyim&Ayakkabı	101	64	324	**60401**	84	42	38	14	1293
Kitap, Müzik, Film, Oyun	755	488	573	142	**59888**	143	116	139	458
Kozmetik&Kişisel Bakım	623	54	561	49	314	**60558**	48	24	113
Mücevher&Saat	74	51	204	75	200	75	**61626**	28	106
Otomotiv&Motosiklet	37	93	153	97	286	76	39	**61242**	484
Spor&Outdoor	259	152	369	1905	645	100	186	589	**58529**

Tablo 5 Karar Ağacı sınıflandırıcısı kullanıldığında Hata Matrisi sonuçları

Veri Seti	Anne&Bebek	Elektronik	Ev&Yaşam	Giyim&Ayakkabı	Kitap, Müzik, Film, Oyun	Kozmetik& Kişisel Bakım	Mücevher&Saat	Otomotiv& Motosiklet	Spor&Outdoor
Anne&Bebek	**59703**	162	598	75	737	719	55	69	250
Elektronik	143	**61132**	290	36	385	67	38	113	214
Ev&Yaşam	608	325	**58681**	349	1002	659	170	177	479
Giyim&Ayakkabı	111	83	352	**60187**	139	48	47	38	1507
Kitap, Müzik, Film, Oyun	839	562	843	159	**58871**	273	167	253	653
Kozmetik&Kişisel Bakım	874	74	804	76	367	**59972**	77	113	167
Mücevher&Saat	74	83	255	60	207	89	**61433**	60	146
Otomotiv&Motosiklet	79	162	225	86	330	91	44	**60826**	790
Spor&Outdoor	244	232	475	1752	748	146	177	653	**58141**

Tablo 6 DVM sınıflandırıcısı kullanıldığında Hata Matrisi sonuçları

Veri Seti	Anne&Bebek	Elektronik	Ev&Yaşam	Giyim&Ayakkabı	Kitap, Müzik, Film, Oyun	Kozmetik& Kişisel Bakım	Mücevher&Saat	Otomotiv& Motosiklet	Spor&Outdoor
Anne&Bebek	**59626**	180	452	79	750	1043	69	31	165
Elektronik	109	**61658**	191	49	241	38	36	67	112
Ev&Yaşam	510	247	**59803**	334	601	462	152	59	170
Giyim&Ayakkabı	104	63	325	**60162**	55	22	54	13	1614
Kitap, Müzik, Film, Oyun	725	289	549	62	**60108**	114	99	104	306
Kozmetik&Kişisel Bakım	650	32	526	25	82	**61133**	78	16	39
Mücevher&Saat	61	44	173	64	124	77	**61793**	16	64
Otomotiv&Motosiklet	39	72	58	30	113	30	29	**61670**	431
Spor&Outdoor	185	148	262	1836	356	33	191	408	**59610**

Tablo 7 MLP (aktivasyon='ReLU') sınıflandırıcısı kullanıldığında Hata Matrisi sonuçları

Veri Seti	Anne&Bebek	Elektronik	Ev&Yaşam	Giyim&Ayakkabı	Kitap, Müzik, Film, Oyun	Kozmetik& Kişisel Bakım	Mücevher&Saat	Otomotiv& Motosiklet	Spor&Outdoor
Anne&Bebek	**59800**	164	564	88	634	974	64	41	182
Elektronik	113	**61478**	216	76	239	43	65	93	134
Ev&Yaşam	400	252	**60158**	295	529	445	156	65	213
Giyim&Ayakkabı	86	54	342	**60246**	43	36	40	31	1503
Kitap, Müzik, Film, Oyun	748	324	622	79	**59989**	112	146	149	407
Kozmetik&Kişisel Bakım	576	57	535	27	116	**61055**	68	28	75
Mücevher&Saat	68	62	192	66	150	114	**61820**	54	100
Otomotiv&Motosiklet	48	54	66	38	145	41	25	**61772**	463
Spor&Outdoor	170	134	271	1528	332	78	159	563	**59012**

Tablo 8 MLP (aktivasyon='logistic sigmoid') sınıflandırıcısı kullanıldığında Hata Matrisi sonuçları

Veri Seti	Anne&Bebek	Elektronik	Ev&Yaşam	Giyim&Ayakkabı	Kitap, Müzik, Film, Oyun	Kozmetik& Kişisel Bakım	Mücevher&Saat	Otomotiv& Motosiklet	Spor&Outdoor
Anne&Bebek	**60104**	138	434	64	636	852	68	26	128
Elektronik	156	**61222**	188	62	251	34	46	71	88
Ev&Yaşam	445	260	**59902**	270	535	456	209	57	119
Giyim&Ayakkabı	77	67	316	**60850**	79	49	66	35	1159
Kitap, Müzik, Film, Oyun	736	297	568	57	**60050**	98	205	132	284
Kozmetik&Kişisel Bakım	756	46	458	59	86	**61082**	83	39	37
Mücevher&Saat	76	54	146	58	140	75	**62102**	16	49
Otomotiv&Motosiklet	22	125	77	34	147	46	55	**61742**	385
Spor&Outdoor	258	167	307	1955	484	53	199	544	**58589**

Tablo 9 Kullanılan sınıflandırıcıların f1-score sonuçları

	Multinominal NB	Lojistik Regression	Random Forest	Karar Ağacı	DVM	MLP (akt.='ReLU')	MLP (akt.='log. sig.')
Anne & Bebek	0.92	0.95	0.96	0.95	0.96	0.96	0.96
Elektronik	0.96	0.98	0.98	0.98	0.98	0.98	0.98
Ev & Yaşam	0.92	0.96	0.95	0.94	0.96	0.96	0.96
Giyim & Ayakkabı	0.94	0.96	0.96	0.96	0.96	0.97	0.97
Kitap, Müzik, Film, Oyun	0.94	0.96	0.95	0.94	0.96	0.95	0.96
Kozmetik & Kişisel Bakım	0.95	0.97	0.97	0.96	0.96	0.97	0.97
Mücevher & Saat	0.98	0.99	0.99	0.99	0.97	0.99	0.99
Otomotiv & Motosiklet	0.97	0.99	0.98	0.97	0.99	0.98	0.99
Spor & Outdoor	0.91	0.95	0.94	0.93	0.99	0.95	0.95
Ortalama	**0.9433**	**0.9678**	**0.9644**	**0.9578**	**0.9700**	**0.9678**	**0.9700**

4. Sonuçlar

Bu çalışmada bir e-ticaret sitesinden alınan tekst verileri birden fazla sınıflandırıcı yöntemi kullanılarak 6 ayrı katagoriye sınıflandırılmış ve çıkan sonuçların doğruluk oranları tablolarla karşılaştırılmıştır. Öncelikle veri önişleme aşamasında kullanılan iki farklı özellik çıkartma işlemi olan TF-IDF ve CountVectorizer karşılaştırması yapılmış ve CountVectorizer'ın daha iyi sonuç verdiği gözlemlenmiştir. Tablo 2'den Tablo 8'e kadar, kullanılan sınıflandırıcıların hata matrisleri gösterilmiştir. Tablolardan görüldüğü üzere Giyim&Ayakkkabı veri seti ile Spor&Outdoor veri seti ve Ev&Yaşam veri seti ile Kitap, Müzik, Film, Oyun veri setinin birbirine yakın olmasından dolayı yanlış sınıflandırma oranı diğer hata oranlarına göre daha yüksek çıkmıştır. Tablo 9'da kullanılan sınıflandırıcıların f1-score değerleri ile katagorilerin ortalama değerleri gösterilmiştir. Tablodan da görüldüğü üzere, kullanılan sınıflandırıcıların ortalama sonuçları kıyaslandığında, DVM ve MLP (aktivasyon='logistic sigmoid') sınıflandırıcısının %97 civarındaki başarım oranları ile diğer yöntemlere göre daha yüksek sonuçlar verdiği görülmektedir.

Kaynaklar

[1] A. H. Aliwy ve E. H. Abdul Ameer, "Comparative Study of Five Text Classification Algorithms with their Improvements", International Journal of Applied Engineering Research, 2017.

[2] H. Alshalabi, S. Tiun, N. Omar, M. Albared, "Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization", Science Direct, Procedia Technology, Elsevier, 2013.

[3] I. Hmeidi, M. Al-Ayyoub, N. A. Abdulla, A. A. Almodawar, R. Abooraig, N. A. Mahyoub, "Automatic Arabic Text Categorisation: A Comprehensive Comparative Study", Journal of Information Science, 2015.

[4] C. C. Aggarwal ve C. X. Zhai, "A Survey Of Text Classification Algorithms", Mining Text Data, Chapter 6, 2012.

[5] H. Deng, Y. Sun, Y. Chang, J. Han, "Probabilistic Models for Classification" C.C. Aggarwal (Eds.), Data Classification Algorithms and Applications (pp. 67-70), CRC Press, New York, USA, 2015.

[6] J. D. Rennie, L. Shih, J. Teevan, D. Karger, "Tackling the poor assumptions of naive bayes text classifiers" Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[7] D. G. Kleinbaum, ve M. Klein, "Logistic Regression: A Self-Learning Text (Statistics for Biology and Health)", Third Edition. New York: Springer 2010.

[8] G. Louppe, "Understanding Random Forest", doktora tezi, University of Liege, 2015.

[9] S. C. Albright, W. L. Winston, ve C. Zappe, "Data Analysis & Decision Making", Üçüncü Baskı, Australia: Thomson South-Western, 2006.

[10] S. R. Gunn, "Support vector machines for classification and regression", Technical Report, Faculty of Engineering, Science and Mathematics, School of Electronics and computer Science, 1998.

[11] J.M. Zurada, "Introduction to Artificial Neural Systems", West Publishing Company, 1992.



Serap Kazan, 1978 yılında Sakarya'da doğmuştur. Lisans öğrenimini 2000 yılında, SAÜ, Elektrir-Elektronik Mühendisliği Bölümünde, yüksek lisans öğrenimini 2003 yılında, SAÜ, Bilgisayar Mühendisliği Bölümünde, doktora öğrenimini 2009 yılında, SAÜ, Elektrik-Elektronik Mühendisliği Bölümünde tamamlamıştır. 2000-2009 yılları arasında SAÜ, Bilgisayar Mühendisliği bölümünde araştırma görevlisi olarak çalışmıştır. 2010'dan beri aynı bölümde öğretim üyesi olarak görev yapmaktadır.

Hakan Karakoca, 1994'de İstanbulda doğmuştur. Lisans öğrenimini 2018 yılında, SAÜ, Bilgisayar Mühendisliği Bölümünde tamamlamıştır.