ESTUDAM Bilişim Dergisi
Cilt **, Sayı * , **-**, 20**

Journal of ESTUDAM Information
Volume **, Issue *, **-**, 20**

DOI: *****

(Geliş Tarihi / Received Date: **.**.20**, Kabul Tarihi/ Accepted Date: **.**.20**)

Türkçe Haber Metinlerinin Makine Öğrenmesi Yöntemleri Kullanılarak Sınıflandırılması

Osman USLU [*1], Serel AKYOL[2]

[1] Eskişehir Osmangazi Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Eskişehir,
ORCID No : https://orcid.org/0000-0002-4377-5952

[2] Eskişehir Osmangazi Üniversitesi, Sivrihisar Meslek Yüksekokulu, Bilgisayar Programcılığı Programı, Eskişehir,
ORCID No : https://orcid.org/0000-0002-5344-4065

Anahtar Kelimeler:
Makine Öğrenmesi,
Metin Sınıflandırma,
Metin Analizi,
Destek Vektör Sınıflandırıcısı,
Rastgele Orman,
Naive Bayes Sınıflandırıcı

Özet:
En büyük bilgi kaynağının internet olarak kabul edildiği günümüz bilgi çağında, elektronik ortamda yer alan metinlerin gün geçtikçe artması sonucunda metin madenciliği ve makine öğrenimi konusu önem kazanmıştır. Teknolojinin gelişmesine paralel olarak bu alanlarda da yenilikler geliştirilmektedir. Yapılan yenilikler ile herhangi bir platformda düzensiz olarak bulunan metinlerin, anlamlı bir bütün haline getirilerek sınıflandırılması ihtiyacı doğmaktadır. Bu çalışmada; farklı makine öğrenmesi yöntemleri kullanılarak Türkçe haber metinlerinin sınıflandırması yapılmaktadır. Haber içerikleri olarak birçok haber metninin ve haber kategorisinin yer aldığı bir veri seti kullanılmıştır. Çalışmada, Destek Vektör Sınıflandırıcısı, Rastgele Orman ve Naive Bayes Sınıflandırıcına göre gerçekleştirilen analiz sonuçları karşılaştırılarak, en başarılı performansa sahip yöntemin 91% doğruluk oranı ile Naive Bayes Sınıflandırıcısı olduğu görülmüştür.

Turkish News Articles Classification Using Machine Learning Techniques

Keywords:
Machine Learning,
Text Classification,
Text Analysis,
Support Vector Classifier,
Random Forest,
Naive Bayes Classifier (NBC)

Abstract: In today's information age, where the largest source of information is accepted as the internet, the issue of text mining and machine learning has become important as a result of the increasing amount of texts in the electronic environment. In parallel with the advancement of technology, innovations are being developed in these areas. Due to the innovations, the need arises to classify the texts found irregularly on any platform into a meaningful whole. In this study; Turkish news texts are classified using different machine learning methods. A data set containing many news texts and news categories was used as news content. In the study, comparing the analysis results performed according to the Support Vector Classifier, Random Forest and Naive Bayes Classifier, it was seen that the method with the most successful performance was the Naive Bayes Classifier with 91% accuracy.

1. GİRİŞ

Yaşadığımız bilgi ve teknoloji çağında, internet üzerinden yapılan paylaşımların artması ve büyük veri setlerinin oluşması nedeniyle aranan veriye doğru bir şekilde ve kısa sürede erişim oldukça önemli bir konu haline gelmiştir [1]. İstenen bilgiye ait veri, görüntü, video, ses ya da metin şeklinde depolanabilmektedir. Bu sonsuz bilgi havuzunda amaçlanan şekilde veriye ulaşım için literatürde veri türüne göre kullanılan farklı algoritma ve yöntemler bulunmaktadır.

Verilerin metin formunda tutulduğu kaynakların çoğunda yapısal ya da yapılandırılmamış formda yer alan birçok içerik bulunmaktadır. Bu içerikler arasında uygun kategorize işlemlerinin gerçekleştirilerek, veri sınıflandırılmasının yapıldığı veriler olabildiği gibi, düzenlenerek her metnin ait olduğu bir başlık ya da sınıf tanımlanmasının yapılması gerektiği yapılandırılmamış formda veriler de olabilmektedir. Bu amaçla bir sınıflandırma yapılmak istenildiğinde veriler hassasiyetle irdelenmeli ve ayrılmak istenen sınıflara göre tasnif işlemi gerçekleştirilmelidir. Literatürde, veri madenciliğinin alt alanı olan, metin sınıflandırması olarak kabul gören bu işlem, bir dokümanda yer alan verilerin özelliklerine bakılarak, verinin önceden tanımlanmış olan kategorilerden hangisine dahil edileceğinin belirlenmesi şeklinde ifade edilmektedir [2].

* İlgili Yazar: uslu_osman@hotmail.com

Bilgi alma, bilgi çıkarımı, bilgi filtreleme, duyarlılık analizi, öneri sistemleri, bilgi yönetimi, metin özetleme gibi metin sınıflandırma uygulamaları belge erişim ve organizasyonu, e-postaların sınıflandırılması ve istenmeyen e-postaların belirlenmesi, haber organizasyonu ve filtrelenmesi gibi birçok alanda yer almaktadır [3, 5].

Güncel bilgi edinme kaynaklarından biri olan haber ajansları, teknolojik gelişmeler sonucu hizmetlerini internet ortamında sunmaktadırlar. Kullanıcılara aktarılan içeriğin sınıflandırılması, uygun etiketler ile sunulması, kullanıcının doğru habere erişiminde oldukça önemlidir. Bu durum da metin formdaki verilerin artmasını buna bağlı olarak otomatik etiketlenmesi gerekliliğini ve metin sınıflandırma ihtiyacını doğurmuştur [6]. Literatürde metin sınıflandırma algoritmalarının haber metinlerinde değerlendirildiği farklı çalışmalar bulunmaktadır.

Toraman vd. çalışmalarında, Türk haber portallarında kullanılmak üzere otomatik metin kategorizasyonu ile yüksek doğruluk derecesinde bir sınıflandırma aracı sunmayı amaçlamışlardır [7]. Bilkent Haber Portalı kullanılarak oluşturulan, farklı özelliklere sahip iki Türkçe test veri setine C4.5, KNN, Naive Bayes ve SVM yöntemleri uygulanarak sonuçları tartışılmıştır. Dört farklı yöntemin sonuçlarının karşılaştırıldığı çalışmada, haber metinlerinin sınıflandırılmasında diğer kök belirleme algoritmalarının da değerlendirilmesi önerilmiştir.

Türkçe dilbilgisi özelliklerini kullanarak web tabanlı haber metinlerinin sınıflandırıldığı çalışmada, sınıflandırıcıda kullanılan özellik vektörünün boyutu ile sınıflandırıcı başarısı arasındaki ilişki irdelenmiş ve boyut azaltılmasına rağmen başarı değerinin düşmediği bir yöntem önerilmiştir [8]. Çalışma sırasında Naive Bayes, SVM, C4.5 ve Rastgele Orman sınıflandırma metotları analiz edilerek, azaltılmış özellik vektörü kullanımında en yüksek başarı oranının Naive Bayes algoritması ile sağlandığı ifade edilmiştir.

Acı ve Çırak, Türkçe haber metinlerini Konvolüsyonel Sinir Ağları ve Word2Vec kullanarak sınıflandırmışlardır [9]. Çalışmada kullanılan, Turkish Text Classification 3600 (TTC-3600) veri seti üzerinde metin sınıflandırması yapılarak, yazarların önceki çalışma sonuçları ile karşılaştırılmıştır. Buna göre, Word2Vec yönteminin, klasik istatistiksel ve makine öğrenmesine dayalı sınıflandırma algoritmalarından daha yüksek performans gösterdiği belirtilmiştir.

Web haberlerinin metin madenciliği ile incelendiği güncel bir çalışmada, turistik alanların çevresel performansını etkileyen faktörlerin metin analizi ile belirlenmesi amaçlanmıştır [10]. Çin'de ciddi çevresel zorluklarla karşı karşıya olan Ulusal 5A Turist Bölgesi için hızlı bir değerlendirme sunması açısından metin madenciliği yöntemleri ile çevrimiçi haber kaynaklarından elde edilen 1.300.000'den fazla kelime kullanılarak bölgenin çevresel performansı değerlendirilmiş ve sonuçları paylaşılmıştır. Haberlerden metin analizi sonucu elde edilmiş çevresel performansı etkileyen faktörler, ilgili araştırmalar ile karşılaştırılarak, çalışma yaklaşımının etkinliği doğrulanmıştır.

Choi vd. ise çalışmalarında, 7.800'den fazla haber izleyicisine uyguladıkları bir anket ile haber kalitesini tahmin etmeyi hedeflemişlerdir [11]. Bu amaçla, 1.500 haber metnine doğal dil işleme, metin madenciliği ve sinir ağı analizleri uygulamışlardır. Çalışma sonuçları, haber metinlerinin gazetecilik değerlerinin, izleyici tarafından derecelendirilen haber kalitesinin dilbilimsel / biçimsel özelliklerinden daha güçlü yordayıcılar olduğunu göstermektedir. Metodolojik olarak, hesaplama ve metinsel yöntemleri geleneksel sosyal bilim yaklaşımıyla bütünleştiren kapsamlı bir çalışma sunulmuştur.

Mukherjee ve Sarkar, metin madenciliği teknikleri kullanarak çevrimiçi gazetelerde yer alan haberleri suça eğilimli alanları belirlemek için analiz etmişlerdir [12]. Bu sayede, polis teşkilatına bildirilmeyen ya da sıradan kişiler tarafından kolayca erişilemeyecek olan suç bilgilerine çevrimiçi yerel ve ulusal gazete haberlerinden ulaşarak suça eğilimli alanları tespit etmeyi amaçlamışlardır. Çalışmada önerilen Naive Bayes sınıflandırıcısı kullanılan model, suç haberlerini filtreleme ve her haber metninden suç yeri çıkarımı olmak üzere iki temel kısımda ele alınmaktadır.

Camilleri vd., metin madenciliği ile çevrimiçi haberleri analiz ederek, depremlere ilişkin içerik oluşturmayı hedeflenmişlerdir [13]. Çevrimiçi haberler ile dünya çapında meydana gelen sismik olaylar arasındaki ilişkinin gerçek zamanlı olarak araştırıldığı çalışmada, deprem ile ilgili raporlardan bilgiler metin madenciliği araçları ile otomatik olarak toplanarak tanımlanmakta ve sınıflandırılmaktadır. Çalışmada, dünyanın farklı yerlerinde bulunan 23 haber ajansı tarafından yayınlanan 268.182 haber ve bültende listelenen büyüklükleri 4 ile 8.2 arasında değişen 14.717 deprem verileri kullanıldığı belirtilmiştir.

Finans haberlerinin kurumsal kredi riski üzerindeki etkisinin irdelendiği çalışmada, metin madenciliği ile elde edilen verilerden lojistik regresyon modeli oluşturulmuştur [14]. Çalışma sonuçlarına göre, erken risk uyarısı oluşturmada, finansal gösterge ve haber metinlerinin yer aldığı, önerilen Lojistik regresyon model doğruluğunun, yalnızca finansal göstergelerle oluşturulan Lojistik regresyon model performansından daha yüksek olduğu gösterilmektedir.

Bilgi işlem kapasitesinin ve veri kaynaklarının artması sonucunda geniş kullanım alanı bulan makine öğrenmesi (machine learning) yöntemleri, literatürde de sunulduğu üzere, her alanda olduğu gibi metin sınıflandırma alanında da öncelikli olarak yer bulmaktadır [15]. Metin sınıflandırma problemi birçok farklı uygulama alanında, farklı makine öğrenme yöntemleri kullanılarak veri setine, model doğruluk ve performansına göre değerlendirilmektedir. Bu çalışmada, veri seti üzerinde temel modelleme denemeleri yapılarak en doğru makine öğrenmesi algoritmasının seçilmesi amaçlanmıştır. Bu amaçla, Türkçe haber metinlerinin yer aldığı bir hazır veri seti için üç farklı makine öğrenmesi yöntemi ile metin

sınıflandırma işlemi gerçekleştirilerek, yöntemlerin performansları karşılaştırılmaktadır.

2. MATERYAL VE METOT

Mevcut bir metnin önceden belirlenen sınıflardan hangisine ya da hangilerine dahil edileceğinin belirlendiği metin sınıflandırma işlemi en temelinde, $T=\{t_1,t_2,\ldots,t_n\}$ kümesinde yer alan her bir metin ya da belgenin, önceden tanımlanmış olan $C=\{c_1,c_2,\ldots,c_m\}$ kümesindeki sınıflara ait olup olmadığının belirlenmesi şeklinde gerçekleştirilmektedir [15]. Buna göre değerlendirilen bir t_i belge ya da metni için C kümesinden herhangi bir sınıfa dahil olup olmaması durumuna göre değer üretilmektedir. Bu amaçla tasarlanmış farklı makine öğrenmesi yöntemleri bulunmaktadır. Kullanılan yönteme göre sınıflandırıcı yöntemin ürettiği sınıflandırma sonuçları ile gerçek sonuçların ne kadar benzerlik gösterdiği belirlenerek, yöntem doğruluğu değerlendirilmektedir.

Bu çalışmada makine öğrenmesi yöntemlerinden olan Destek Vektör Sınıflandırıcısı (Support Vector Classifier-SVC), Rastgele Orman (Random Forest) ve Naive Bayes yöntemleri ile Türkçe haber metinlerinin sınıflandırma analizleri yapılmaktadır.

Destek Vektör Sınıflandırıcısı (Support Vector Classifier-SVC)

Vapnik tarafından geliştirilmiş destek vektör makinesi yönteminin alt yöntemlerinden biri olan, Destek Vektör Sınıflandırıcıları, doğrusal olmayan karar sınırlarına izin vermesi sebebiyle sınıflandırma işlemlerinde tercih edilmektedir [16]. Sınıflandırma yapılırken temel amaç, sınıfları birbirinden ayıracak olan optimal ayırma hiper düzleminin elde edilmesidir. Bu sayede, farklı sınıflara ait destek vektörleri arasındaki uzaklık maksimize edilmektedir [17]. Destek vektör makinesi temelli yöntemlerin işleyiş sürecinde çekirdek fonksiyonu seçimi ve parametre optimizasyonu oldukça önemli bir konudur. SVC, sadece doğrusal çekirdek yapısını desteklemesi nedeniyle sınıflandırmada daha hızlı bir yöntem olarak öne çıkmaktadır.

Rastgele Orman (Random Forest) Sınıflandırıcısı

Rastgele Orman (Random Forest-RF) Sınıflandırıcısı, birçok karar ağacının biraya gelmesiyle oluşmakta ve ağacı meydana getiren bireysel ağaçlar arasından doğruluğu en yüksek olanlar tercih edilmektedir. Yapıda bulunan ağaçların dalları veri setindeki özelliklere göre oluşmaktadır [18]. Bir sınıflandırıcı yerine birden çok sınıflandırıcı üreten ve sonrasında bu sınıflandırıcıların tahminlerinden alınan oylar ile yeni veriyi sınıflandıran öğrenme algoritmasıdır [19]. Dallanma bu özelliklerde karar noktalarına bağlıdır [20]. Karar ağacı modellerinin en önemli dezavantajı ise küçük veri setlerinde yaşanan aşırı uyum durumudur. RF, analiz edilen örneklem boyutu ile alt örnek boyutunun aynı olduğu, doğruluk performansının geliştirilmesi ve aşırı uyumun kontrol edilmesi için ortalamayı kullanan bir meta tahmin sınıflandırıcısıdır.

Naive Bayes Sınıflandırıcısı

Naive Bayes, basit kullanımı ve etkinliği nedeniyle literatürde pek çok metin sınıflandırma uygulaması bulunan, pratik bir istatistiksel sınıflandırıcıdır [21]. Basit yapısına karşın hesaplama ve doğru sınıflandırma oranı yüksek bir sınıflandırıcı olması sebebiyle metin madenciliğinin de dahil olduğu pek çok alanda tercih edilmektedir [5].

Bayes teoremine göre metin sınıflandırması yapılırken, d_j belgesinin bir c sınıfına ait olma olasılığı şu şekilde hesaplanmaktadır [16, 3]:

$$p(c|d_j) = \frac{p(d_j|c)p(c)}{p(d_j)} = \frac{p(d_j|c)p(c)}{p(d_j|c)p(c) + p(d_j|\bar{c})p(\bar{c})}$$

$$p(c|d_j) = \frac{\frac{p(d_j|c)}{p(d_j|\bar{c})} \cdot p(c)}{\frac{p(d_j|c)}{p(d_j|\bar{c})} \cdot p(c) + p(c)} \qquad (1)$$

Veri Seti

Çalışmada makine öğrenmesi yöntemlerinin analizlerinin uygulandığı veri seti, metin formda, iki sütun ve 4900 satırdan oluşan Türkçe haber metinlerini içermektedir [22]. Veri setinde sütunlardan ilki kategori başlığını (sınıf bilgisi) ifade ederken, ikinci sütunda metin başlığı altındaki veriler yer almaktadır. Haber metinleri, spor, dünya, ekonomi, kültür, teknoloji, siyaset ve sağlık olmak üzere 7 kategoriye ayrılmaktadır. Her bir kategoride homojen ve eşit olarak dağılmış olan 700 adet veri yer almaktadır.

Çalışmada, Türkçe haber metin içeriklerinin sınıflandırılmasında; nesne yönelimli, yorumlamalı, birimsel ve etkileşimli yüksek seviyeli bir programlama dili olan Pyton programlama dili ve GPU'lara açık erişim imkanı sunan, Google Colaboratory- "Colab" platformu kullanılmıştır.

Uygulamada Numpy, Pandas, Matplotlib ve Searborn Python Kütüphanelerinden ve Doğal Dil Araç Setinden faydalanılmıştır [23].

Numpy: (Numerical Python) Bilimsel hesaplamalarda kullanılan temel pakettir. Python'daki çok boyutlu diziler ve matrisler üzerinde yapılan işlemlerde çok sayıda kullanışlı özellikler sunmaktadır. Matematiksel işlemlerde kolaylık sağlayarak hesaplamaların daha hızlı bir şekilde gerçekleşmesini sağlamaktadır.

Pandas: Etiketli ve ilişkisel verilerle basit ve sezgisel olarak çalışmak üzere tasarlanmış Pyhon paketidir. Hızlı ve kolay bir şekilde veri işleme, veriyi yönetme, görselleştirme için tasarlanmış, veri üzerinde ön işleme ve analiz gerçekleştirmek için geliştirilmiş bir araçtır.

Matplotlib: İki boyutlu grafik çizimleri elde etmek için kullanılan bir kütüphanedir. Çeşitli formatlarda ya da

interaktif ortamlarda çıktılar elde etmek için kullanılmaktadır.

Searborn: Ağırlıklı olarak istatistiksel modellerin görselleştirilmesinde kullanılmaktadır. Verileri özetleyen, genel dağılımları gösteren görselleştirmeler sunmaktadır. Temel olarak Matplotlib'e dayanmaktadır.

Natural Language Toolkit (NLTK) [24] - Doğal Dil Araç Seti; Python programlama dilinde yazılmış öncelikle İngilizce için sembolik ve istatistiksel doğal dil işleme için kullanılan kütüphaneler ve programların yer aldığı pakettir. Çalışmada analiz edilen verilerin Türkçe olması sebebiyle, söz konusu paketin Türkçe seti kullanılmaktadır. Uygulamada NLTK sayesinde Türkçe durdurma kelimeleri (stopwords) eklenilerek, gereksiz kelimeler listelenmiş ve metinlerden ayrıştırılmıştır.

Yazılımda, değişken tanımlamaları yapılarak kategori ve metinlerin değişkenlere ataması yapılmıştır. Y değişkeni kategorileri temsil ederken, işlenmesi hedeflenen veriler X değişkenine atanmıştır.

```
y = data.category.values
x = data.text.values
```

Veri ön işleme sayesinde, analiz edilecek verilerin, kullanılacak yönteme uygun veri haline getirilmesi sağlanmaktadır. Bu amaçla, Python'da CountVectorizer nesnesi kullanılarak, X değişkenine atanmış olan metinler matrislere çevrilerek çalışmada kullanılacak başka bir forma dönüştürülmüştür.

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer=CountVectorizer(min_df=5,stop_words=stop,ngram_range=(1,3))
vectorizer.fit(x)

CountVectorizer(analyzer='word', binary=False, decode_error='strict', dtype=<class 'numpy.int64'>, encoding='utf-8', input='content', lowercase=True, max_df=1.0, max_features=None, min_df=5,ngram_range=(1, 3), preprocessor=None, stop_words=['acaba', 'ama', 'aslında', 'az', 'bazı', 'belki', 'biri', 'birkaç', 'birşey', 'biz', 'bu', 'çok','çünkü', 'da', 'daha', 'de', 'defa', 'diye', 'eğer','en', 'gibi', 'hem', 'hep', 'hepsi', 'her', 'hiç','için', 'ile', 'ise', 'kez', ...], strip_accents=None,token_pattern='(?u)\\b\\w\\w+\\b',tokenizer=None,vocabulary=None)
```

Veri setinde modelleme denemelerinin yapılacağı örneklemler rastgele belirlenerek form dönüşümleri yapılmıştır. Sonraki aşama olarak, belirlenen makine öğrenmesi algoritmalarına bu dönüşüm verileri gönderilerek işlenmesi sağlanmış ve sınıflandırma işlemleri gerçekleştirilmiştir. Buna göre üç farklı makine öğrenmesi yöntemi ile oluşturulan modellerin başarı metrikler Tablo 1-2 ve 3'de gösterilmiştir.

Tablo 1. SVC Model Sonuçları

	precision	recall	f1-score	support
dünya	0.84	0.79	0.82	174
ekonomi	0.83	0.88	0.85	163
kültür	0.94	0.90	0.92	181
sağlık	0.96	0.93	0.94	156
siyaset	0.84	0.92	0.88	181
spor	0.93	0.95	0.94	186
teknoloji	0.90	0.86	0.88	184
accuracy			0.89	1225
macro avg	0.89	0.89	0.89	1225
weighted avg	0.89	0.89	0.89	1225

Tablo 2. RF Model Sonuçları

	precision	recall	f1-score	support
dünya	0.84	0.75	0.79	184
ekonomi	0.76	0.87	0.81	152
kültür	0.94	0.83	0.88	196
sağlık	0.96	0.90	0.93	162
siyaset	0.85	0.88	0.87	192
spor	0.88	0.98	0.93	171
teknoloji	0.84	0.88	0.86	168
accuracy			0.87	1225
macro avg	0.87	0.87	0.87	1225
weighted avg	0.87	0.87	0.87	1225

Tablo 3. NB Model Sonuçları

	precision	recall	f1-score	support
dünya	0.83	0.86	0.84	158
ekonomi	0.84	0.87	0.86	167
kültür	0.94	0.91	0.93	179
sağlık	0.96	0.95	0.96	152
siyaset	0.91	0.90	0.90	201
spor	0.95	0.99	0.97	182
teknoloji	0.92	0.87	0.90	186
accuracy			0.91	1225
macro avg	0.91	0.91	0.91	1225
weighted avg	0.91	0.91	0.91	1225

3. BULGULAR

Veri analizinde, en doğru modelin hangisi olması gerektiğine karar vermek için model çıktılarının dikkatle değerlendirilmesi gerekmektedir. Literatürde kabul görmüş, model performansı, model etkinliğini belirlemede kullanılan çeşitli yöntemler bulunmaktadır.

Bu yöntemlerden en sık kullanılanları, Kesinlik (Precision), Duyarlılık (Recall) ve Doğruluk (Accuracy) metrikleridir [21]. Bu değerlerin belirlenmesinde öncelikle modelin, TP (doğru pozitif) ve TN (doğru negatif) doğru sınıflandırmaları; FP (yanlış pozitif) ve FN (yanlış negatif) yanlış sınıflandırmaları gösteren ifadeler analiz edilmektedir.

Kesinlik (π_i), herhangi bir *d* belgesinin, c_i sınıfına dahil edilmesi durumunda, bu sınıflandırmanın doğru olma olasılığını göstermektedir [21].

$$\pi_i = \frac{TP_i}{TP_i + FP_i} \quad (2)$$

Duyarlılık (p_i), gerçekte c_i sınıfı altında bulunması gereken belgelerin kaçının bu sınıfta yer aldığı şeklinde tanımlanmaktadır [15, 21].

$$p_i = \frac{TP_i}{TP_i + FN_i} \tag{3}$$

Doğruluk (A_i), sınıflandırıcının doğru sonuçlar elde etmekteki yeteneğini göstermekte ve eşitlik 4'teki gibi hesaplanmaktadır [21]. Hata oranı ise bu değerin 1'e tamlayanı olarak bulunmaktadır.

$$A_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{4}$$

F1-skoru, uç durumların göz ardı edilmesini engellemek için Kesinlik (Precision) ve Duyarlılık (Recall) değerlerinin harmonik ortalaması ifade etmektedir.

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

Aynı veri seti için, SVC, Random Forest ve Naive Bayes algoritmaları ayrı ayrı değerlendirilerek sınıflandırma işlemi başarıyla gerçekleştirilmiştir. Yöntemlerin sınıflandırma başarı oranları farklılık gösterse de, her bir algoritmanın sınıflandırma konusunda yeterli olduğu görülmüştür.

Uygulama çıktılarına göre, sadece doğruluk değerlerine göre model seçimi yapılması yeterli değildir. Yapılan analiz sonuçlarına göre SVC algoritması; 89% F1 skoru elde ederek çalışmada 2.sırada yer almıştır. Random Forest algoritması; yeterli olarak kabul edilse de 87% oranında F1 skoru ile çalışmada 3.algoritma olarak yer almıştır. Naive Bayes algoritması ise; 91% F1 skoru değeri ile en başarılı algoritma olmuştur.

4. TARTIŞMA VE SONUÇ

En büyük bilgi kaynağının internet olarak kabul edildiği günümüz bilgi çağında, aranılan bilgiye erişimin hızlı ve doğru bir şekilde ve bilginin bulunabilirliğinin kolay olması gerekmektedir. Bu noktada metin madenciliği yöntemlerinden faydalanılmaktadır. Türkçe haber metinlerinin analiz edildiği bu çalışmada, literatürden farklı olarak, yüksek doğruluk derecelerine sahip, farklı makine öğrenmesi yöntemleri ile karşılaştırılmış 3 yöntem birlikte incelenmiştir. Makine öğrenmesi yöntemlerinden SVC, RF ve NB yöntemleri ile metin sınıflandırma işleminde tüm yöntemler başarıya ulaşmış ve seçilen tüm yöntemlerin model başarı oranları oldukça yüksek bulunmuştur. Sonuç olarak, Türkçe haber metinlerinin sınıflandırılmasında en başarılı performansı Naive Bayes (91%) yöntemi göstermiştir.

Çalışmanın sonuçlarının derin öğrenme, yapay sinir ağları gibi farklı yöntemler ile kıyaslanarak hibrit yöntemlerin oluşturulması sonraki çalışmalar için kaynak teşkil edecektir.

Teşekkür

"Bilgi Erişim Sistemleri" dersi projesinden hareketle geliştirilen bu çalışma için değerli Prof. Dr. Eyyüp GÜLBANDILAR'a teşvik ve desteklerinden ötürü teşekkür ederiz.

KAYNAKÇA

[1] D. Kılınç, E. Borandağ, F. Yücalar, V. Tunali, M. Şimşek, ve A. Özçift. 2016. KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi," Marmara Fen Bilim. Derg., 28(3), 89–94.

[2] H. K. Yıldız, M. Genctav, N. Usta, B. Diri, ve M. F. Amasyali. 2007. Metin Sınıflandırmada Yeni Özellik Çıkarımı. 2007 IEEE 15th Signal Processing and Communications Applications, 1–4.

[3] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown. 2019. Text Classification Algorithms: A Survey. Information, 10(4), 1–68.

[4] C. C. Aggarwal , C. Zhai. 2012. Mining Text Data. Springer, New York.

[5] A. Onan ve S. Korukoğlu. 2016. Metin sınıflandırmada öznitelik seçim yöntemlerinin değerlendirilmesi. Akademik Bilişim.

[6] S. Yıldırım ve T. Yıldız. 2018. Türkçe için Karşılaştırmalı Metin Sınıflandırma Analizi. Pamukkale Üniversitesi Mühendislik Bilim. Derg., 24(5), 879–886.

[7] C. Toraman, F. Can, S. Koçberber. 2011. Developing a Text Categorization Template for Turkish News Portals. 2011 International Symposium on Inovations in Intelligent Systems and Applications, 379–383.

[8] P. Tüfekci, E. Uzun, ve B. Sevinç. 2012. Türkçe Dilbilgisi Özelliklerini Kullanarak Web Tabanlı Haber Metinlerinin Sınıflandırılması. 2012 20th Signal Processing and Communications Applications Conference (SIU), 1–4.

[9] Ç. İ. Acı ve A. Çırak. 2019. Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması. Bilişim Teknol. Derg., 12(3), 219–228.

[10] F. Wang, X. Peng, Y. Qin, C. Wang. 2020. What can the news tell us about the environmental performance of tourist areas? A text mining approach to China's National 5A Tourist Areas. Sustain. Cities Soc., 52(101818).

[11] S. Choi, H. Shin, S. S. Kang. 2020. Predicting Audience-Rated News Quality: Using Survey, Text Mining, and Neural Network Methods. Digit. Journal., 1–22.

[12] S. Mukherjee, K. Sarkar. 2020. Analyzing Large News Corpus Using Text Mining Techniques for Recognizing High Crime Prone Areas. 2020 IEEE Calcutta Conference, CALCON, 444–450.

[13] S. Camilleri, M. R. Agius, J. Azzopardi. 2020. Analysis of Online News Coverage on Earthquakes Through Text Mining. Front. Earth Sci., 8(May), 1–12.

[14] C. Li, Q. Liu, L. Huang. 2020. Credit Risk Management of Scientific and Technological Enterprises Based on Text Mining. Enterp. Inf. Syst., 1–17.

[15] A. C. Tantuğ.2012. Metin Sınıflandırma (Text Classification). Türkiye Bilişim Vakfı Bilgi. Bilim. ve Mühendisliği Derg., 5(2).

[16] G. S. Chavan, S. Manjare, P. Hegde, A. Sankhe. 2014. A Survey of Various Machine Learning Techniques for Text Classification. Int. J. Eng. Trends Technol., 15(6), 288–292.

[17] S. Ayhan ve Ş. Erdoğmuş. 2014. Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi. Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilim. Derg., 9(1), 175–201.

[18] Ü. Veranyurt, A. F. Deveci, M. F. Esen, ve O. Veranyurt. 2020. Makine Öğrenmesi Teknikleriyle Hastalık Sınıflandırması: Random Forest, K-Nearest Neighbour Ve Adaboost Algoritmaları Uygulaması. Uuslararası Sağlık Yönetimi ve Strat. Araştırma Derg., 6(2), 275–286.

[19] M. Bilgin. 2017. Gerçek Veri Setlerinde Klasik Makine Öğrenmesi Yöntemlerinin Performans Analizi. Breast, 2(9), 683–688.

[20] L. Breiman. 2001. Random Forest. Mach. Learn, 45, 5–32.

[21] M. Ikonomakis, S. Kotsiantis, V. Tampakas. 2005. Text Classification Using Machine Learning Techniques. WSEAS Trans. Comput., 4(8), 966–974.

[22] S. Yıldırım. 2017. Text Categorization for Turkish - Multi NB. https://www.kaggle.com/savasy/text-categorization-for-turkish-multi-nb. (Erişim Tarihi: 28.11.2020).

[23] A. Yeşilyurt. 2018. Veri Bilimi için Python Kütüphaneleri. https://medium.com/@amine.yesilyurt/python-kutuphaneleri-e59fe08cc276. (Erişim Tarihi: 29.11.2020)

[24] Wikipedia. Natural Language Toolkit. https://en.wikipedia.org/wiki/Natural_Language_Toolkit . (Erişim Tarihi: 29.11.2020)