İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ
2010
FEN BİLİMLERİ ENSTİTÜSÜ

Haber Türünün Belirlenmesinde Makine Öğrenmesi Yöntemlerinin Performans Analizi

Yazılım Mühendisliği Ana Bilim Dalı

Yüksek Lisans Bitirme Projesi

Güzel Başpınar

Bitirme Proje Danışmanı: Prof. Dr. Ayşegül Alaybeyoğlu Soy

Haziran 2023

İzmir Kâtip Çelebi Üniversitesi Fen Bilimleri Enstitüsü öğrencisi **Güzel Başpınar** tarafından hazırlanan **Haber Türünün Belirlenmesinde Makine Öğrenmesi Yöntemlerinin Performans Analizi** başlıklı bu çalışma tarafımızca okunmuş olup, yapılan savunma sınavı sonucunda kapsam ve nitelik açısından başarılı bulunarak jürimiz tarafından YÜKSEK LİSANS BİTİRME PROJESİ olarak kabul edilmiştir.

ONAYLAYANLAR:

Tez Danışmanı: **Prof. Dr. Ayşegül Alaybeyoğlu Soy**
İzmir Kâtip Çelebi Üniversitesi

Yazarlık Beyanı

Ben, **Güzel Başpınar**, başlığı **Haber Türünün Belirlenmesinde Makine Öğrenmesi Yöntemlerinin Performans Analizi** olan bu bitirme projesi ve bitirme projesi içinde sunulan bilgilerin şahsıma ait olduğunu beyan ederim. Ayrıca:

- Bu çalışmanın bütünü veya esası bu üniversitede Yüksek Lisans derecesi elde etmek üzere çalıştığım süre içinde gerçekleştirilmiştir.
- Daha önce bu bitirme projesinin herhangi bir kısmı başka bir derece veya yeterlik almak üzere bu üniversiteye veya başka bir kuruma sunulduysa bu açık biçimde ifade edilmiştir.
- Başkalarının yayımlanmış çalışmalarına başvurduğum durumlarda bu çalışmalara açık biçimde atıfta bulundum.
- Başkalarının çalışmalarından alıntıladığımda kaynağı her zaman belirttim. Bitirme projesinin bu alıntılar dışında kalan kısmı tümüyle benim kendi çalışmamdır.
- Kayda değer yardım aldığım bütün kaynaklara teşekkür ettim.
- Bitirme projesinde başkalarıyla birlikte gerçekleştirilen çalışmalar varsa onların katkısını ve kendi yaptıklarımı tam olarak açıkladım.

Tarih: 25.06.2023

_____

Haber Türünün Belirlenmesinde Makine Öğrenmesi Yöntemlerinin Performans Analizi

Öz

İnternetin en büyük bilgi kaynağı olarak kabul edildiği bilgi çağında, elektronik ortamdaki metinlerin sayısının giderek artması, metin madenciliği ve makine öğrenimi konularının önemini arttırmıştır. Teknolojinin gelişmesiyle birlikte, bu alanlarda da sürekli olarak yeni gelişmeler olmaktadır. Geliştirilen yeniliklerle, herhangi bir platformda düzensiz olarak bulunan metinlerin anlamlı bir şekilde birleştirilerek sınıflandırılması gerekliliği ortaya çıkmaktadır.

Bu çalışmada, farklı makine öğrenimi yöntemi kullanılarak haber metinleri sınıflandırıldı ve makine öğrenimi tekniklerinin performansı incelendi. Veri seti olarak kategorilere ayrılmış haber metinleri kullanıldı.

Anahtar Sözcükler: Yapay zeka, makine öğrenmesi, sınıflandırma, metin sınıflandırma

Performance Analysis of Machine Learning Methods in Determining News Genre

Abstract

In the information age where the Internet is considered the largest source of information, the increasing number of texts in electronic media has emphasized the importance of text mining and machine learning. With the advancement of technology, there are constantly new developments in these fields. The innovations that have been developed necessitate the meaningful integration and classification of texts that are irregularly found on any platform.

In this study, news texts were classified using different machine learning methods, and the performance of machine learning techniques was examined. Categorized news texts were used as the dataset.

Keywords: Artificial intelligence, machine learning, classification, text classification

İçindekiler

Şekiller Listesi

Tablolar Listesi

Matematiksel İfadeler Listesi

Kısaltmalar Listesi

NLTK	Natural Language Toolkit
SVM	Support Vector Machine
NLP	Natural Language Processing
AI	Artificial Intelligence
KNN	K-Nearest Neighbors
TF-IDF	Term Frequency–Inverse Document Frequency

Bölüm 1

Giriş

Haber türünün belirlenmesi, haber içeriklerinin otomatik olarak sınıflandırılması ve kategorize edilmesi işlemidir. Bu konu, haber metinlerinin büyük hacimlerini yönetmek ve haberlerin hızla analiz edilmesini sağlamak için önemli bir öneme sahiptir [1]. Haber türünün doğru bir şekilde belirlenmesi, haberlerin doğru hedef kitleye iletilmesi, özel ilgi alanlarına göre filtrelenmesi ve haber akışının düzenlenmesi gibi birçok uygulamada önemli bir rol oynamaktadır.

Son yıllarda, makine öğrenmesi yöntemleri haber türü belirleme problemi için başarılı bir çözüm sunmaktadır. Makine öğrenmesi, büyük veri setleri üzerinde örüntüleri tanımlamak, bilgi çıkarmak ve sınıflandırma yapmak için istatistiksel modeller kullanır [2]. Bu yöntemler, haberlerin içeriğini analiz ederek onları farklı türlerine göre sınıflandırabilen etkili ve esnek bir yaklaşım sağlamaktadır.

Makine öğrenmesi ilk kez 1959'da Arthur Samuel tarafından bir dama oyununu analiz etmek için kullanılmıştır. Arthur Samuel, makine öğrenmesini bilgisayarlara açıkça programlanmadan öğrenme yeteneği veren araştırma alanı olarak tanımlamıştır [6]. 1998 yılında Tom Mitchell tarafından yapılan tanıma göre, "Bir bilgisayar programının, T ile ölçülen performansı P ile ölçüldüğünde iyileşirse, bazı T görevlerine göre E deneyiminden ve bazı performans ölçülerinden P öğrendiği söylenir." [7]

Literatürde, haber türü belirleme üzerine birçok çalışma bulunmaktadır. Önceki araştırmalar, çeşitli makine öğrenmesi algoritmalarının haber türü belirleme problemi üzerinde başarılı sonuçlar elde ettiğini göstermiştir. Bunlar arasında destek vektör makineleri (SVM), karar ağaçları, rastgele ormanlar, derin öğrenme teknikleri gibi yöntemler yer almaktadır [3].

Wang ve arkadaşları (2019) [4], doğal dil işleme ve makine öğrenmesi yöntemlerini kullanarak haber türü belirleme problemine odaklanan bir çalışmayı sunmaktadır. Bu

çalışma, haberlerin içeriğini temsil etmek için kelime dağılımları ve kelime vektörleri gibi özelliklerin kullanıldığı bir sınıflandırma modeli geliştirmiştir. Ayrıca, Smith ve Johnson (2020) [5], haber türü belirleme için destek vektör makinelerini kullanan bir yöntem önermiştir. Bu çalışmada, haberlerin metinsel özellikleri temsil eden kelime ve n-gram özellikleri kullanılarak haberler sınıflandırılmıştır.

Bir diğer ilgi çekici çalışma ise Chen ve arkadaşları tarafından yapılmıştır (2018) [3]. Bu çalışmada, haberlerin türlerini belirlemek için derin öğrenme yaklaşımı kullanılmıştır. Haber metinlerinin özniteliklerini temsil etmek için bir derin sinir ağı modeli kullanılarak haberler sınıflandırılmıştır. Bu çalışma, derin öğrenme yöntemlerinin haber türü belirleme probleminde etkili bir şekilde kullanılabileceğini göstermiştir.

Makine öğrenmesi algoritmalarının haber türü belirleme üzerindeki performansını analiz etmek için çeşitli metrikler kullanılmaktadır. Bu metrikler arasında doğruluk, hassasiyet, özgüllük ve geri çağırma gibi değerlendirme ölçütleri bulunmaktadır. Bu ölçütler, bir sınıflandırma modelinin ne kadar doğru ve hassas bir şekilde haber türlerini tahmin edebildiğini belirlemek için kullanılır.

Bu bitirme projesi, haber türünün belirlenmesi için makine öğrenmesi yöntemlerinin performansını analiz etmeyi amaçlamaktadır. Çeşitli makine öğrenmesi algoritmalarının haber metinlerinin sınıflandırılmasındaki etkinliğini ve doğruluğunu değerlendirecektir.

Bölüm 2

Makine Öğrenimi

Makine öğrenimi, olasılık, istatistik ve optimizasyon gibi matematiksel alt alanlara dayanan bir yaklaşımdır. Bu yöntemde, bir model oluşturulur ve makine, geçmiş verilere dayanarak geleceği tahmin etme yeteneği kazanır.

Makine öğrenimi, büyük veri kümelerindeki kalıpları ve ilişkileri bulmak ve bu analizlere dayalı olarak en iyi kararları ve tahminleri yapmak için kullanılan algoritmaları içerir. Makine öğrenimi uygulamaları, veri miktarı arttıkça daha doğru sonuçlar üretebilme yetenekleriyle birlikte gelişir. Ne kadar çok veriye sahip olunursa, makine öğrenimi modelleri o kadar iyi eğitilir ve daha kesin tahminler yapabilir [8].

"Makine öğrenimi algoritması, sırasıyla öğrenme, test ve uygulama süreçlerinden geçmektedir. Öğrenme, algoritmanın eğitimine; test süreci, gerçek uygulama öncesi algoritmanın tekrar eğitime ihtiyaç duyup duymadığının değerlendirilmesine ve nihayetinde uygulama süreci de, eğitilmiş algoritmanın artık pratikte de kullanılmasına karşılık gelmektedir. Bütün bu mekanizma, bizleri otonom zeki sistemlere götürmektedir" [9].

2.1 Makine Öğrenimi Algoritma Türleri

Makine öğrenme algoritmaları, girdi ve beklenen çıktı türüne bağlı olarak dört farklı öğrenme sitili altında gruplandırılabilir: denetimli öğrenme, denetimsiz öğrenme, takviyeli öğrenme ve yarı denetimli öğrenme.

2.1.1 Denetimli Öğrenme (Supervised Learning)

Denetimli öğrenme (Supervised Learning) yaklaşımında, bir algoritma, bilinen problem verileri ve bu verilere karşılık gelen sonuçlarla eğitilir. Bu veri seti, giriş ve çıkış değerleri arasındaki ilişkiyi yansıtan bir eşleşme fonksiyonu oluşturur.

Denetimli öğrenme yöntemiyle, algoritma, bu eşleşme fonksiyonunu kullanarak yeni giriş verilerini analiz eder ve tahminler yapar. Bu şekilde, öğrenme işlemi gerçekleştirilir ve gelecekteki giriş verilerine dayalı olarak çıktılar tahmin edilebilir [10].

2.1.2 Denetimsiz Öğrenme (Unsupervised Learning)

Denetimsiz öğrenme (Unsupervised Learning) yaklaşımında, bir eğitim veri seti bulunur. Ancak, bu yöntemde elde edilecek sonuçlar önceden tahmin edilemez ve bilinmez. Algoritma, veriler üzerinden geçerken sınıflandırmayı kendi kendine yapar, çıktı verilerine ihtiyaç duymadan öğrenme işlemini gerçekleştirir [10]. Bu yaklaşımda, algoritma, verilerdeki yapıları, desenleri ve ilişkileri keşfetmeye çalışır. Böylece, verilerin gizli özelliklerini belirleyebilir, gruplamalar yapabilir veya veriye yönelik farklı analizler gerçekleştirebilir. Denetimsiz öğrenme, özellikle veri setinin özellikleri ve yapısı hakkında önceden bilgi sahibi olmadığımız durumlarda kullanışlıdır [10]. Genellikle örüntüler arasında benzerliklerden faydalanarak kümeleme yöntemini kullanırlar [11].

2.1.3 Takviyeli Öğrenme (Reinforcement Learning)

Takviyeli öğrenme (Reinforcement Learning) yaklaşımında, algoritma, bir çözümü uyguladığında elde edilen geri bildirimlere dayanarak öğrenme yapar. Bu geri bildirimler, çözümün iyi mi, kötü mü, doğru mu, yanlış mı olduğunu belirtir. Algoritma, bu geri bildirimler sayesinde kendisini daha iyi eğitir ve öğrenir.

Takviyeli öğrenme, bir öğrencinin deneyerek ve geribildirim alarak bir görevi öğrenmesiyle benzerlik gösterir. Algoritma, bir ortamla etkileşime girer, belirli bir eylem gerçekleştirir ve ardından ortamdan bir geri bildirim alır. Bu geri bildirim, genellikle bir ödül veya ceza şeklinde olabilir. Algoritma, elde ettiği geri bildirimlere dayanarak gelecekteki eylemlerini belirlemek için bir strateji geliştirir.

Takviyeli öğrenme, özellikle karmaşık ve dinamik ortamlarda optimal kararlar almayı öğrenmek için kullanılır. Örnek olarak, oyunlarda yapay zeka eğitimi, robot kontrolünde davranış öğrenme veya finansal algoritmaların optimize edilmesi gibi alanlarda takviyeli öğrenme yöntemleri kullanılabilir.

2.1.4 Yarı Denetimli Öğrenme (Semi-Supervised Learning)

Yarı Denetimli Öğrenme (Semi-Supervised Learning) yaklaşımı, hem denetimli öğrenme hem de denetimsiz öğrenme yaklaşımlarının birleşimi olan bir öğrenme modelidir. Yarı denetimli öğrenmede, eğitim veri setinin bir kısmı etiketli verilerden oluşurken, diğer kısmı etiketlenmemiş verilerden oluşur. Etiketli veriler, giriş verilerine karşılık gelen doğru çıktı etiketlerine sahipken, etiketlenmemiş verilerin çıktı etiketleri bilinmemektedir.

Bu yaklaşımda, etiketli verilerin yanı sıra etiketlenmemiş veriler de kullanılarak bir model oluşturulur. Etiketli veriler, modelin doğru sonuçları tahmin etmesini sağlamak için kullanılırken, etiketlenmemiş veriler, veri setindeki genel yapının veya desenlerin öğrenilmesine yardımcı olur.

Yarı denetimli öğrenme, etiketleme maliyetinin yüksek olduğu durumlarda veya sınırlı sayıda etiketli veriye sahip olduğumuzda kullanışlıdır. Etiketlenmemiş verilerin kullanılması, modelin genel performansını artırabilir ve daha iyi bir öğrenme sağlayabilir. Bu yöntem, veri setinin tam potansiyelini değerlendirebilmek ve daha genellemeler yapabilmek için kullanılan bir yaklaşımdır.

2.2 Sınıflandırma Algoritmaları

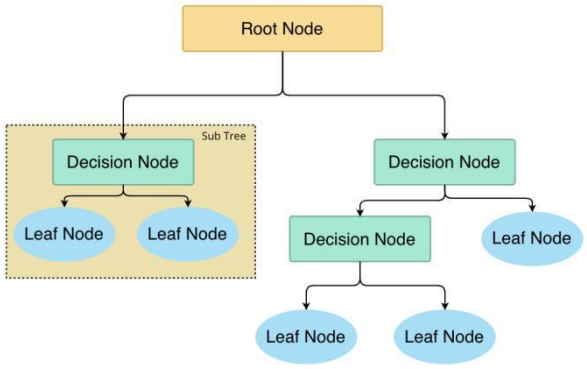
2.2.1 Lojistik Regresyon

Lojistik Regresyon, olasılık kavramına dayanan en büyük olabilirlik kestirimi kullanan makine öğrenimi algoritmasıdır. Regresyon analizi, bir ya da daha fazla bağımsız değişken ile hedef değişken arasındaki ilişkiyi matematiksel olarak modelleyen bir yöntemdir. Lojistik regresyonda sigmoid fonksiyonu kullanılarak kesikli bir değerler alan hedef değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilmektedir. [12] [13] [14]

$$\mathrm{P}(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}} \tag{2.1}$$

2.2.2 Karar Ağaçları

Sınıflama, özellik ve hedefe göre karar düğümleri (decision nodes) ve yaprak düğümlerinden (leaf nodes) oluşan ağaç yapısı formunda bir model oluşturan bir sınıflandırma yöntemidir. Karar ağacı algoritması, veri setini küçük ve hatta daha küçük parçalara bölerek geliştirilir. Bir karar düğümü bir veya birden fazla dallanma içerebilir. İlk düğüme kök düğüm (root node) denir. Bir karar ağacı hem kategorik hem de sayısal verilerden oluşabilir. Birden fazla ayırma yöntemi vardır. Bunlar, sınıflandırma yapmak için belirli kelime varlığını veya yokluğunu arayan tek öz nitelik bölme ve belgedeki kelimeleri önceden tanımlanmış kelimelerle eşleştiren benzerlik tabanlı çoklu öz nitelik bölmedir. Algoritma çalışırken ilk düğümden başlanarak sorular sorulmaya başlanır ve son eleman olan yapraklara ulaşana kadar ağacın büyümesi ve dallanması devam eder [15].



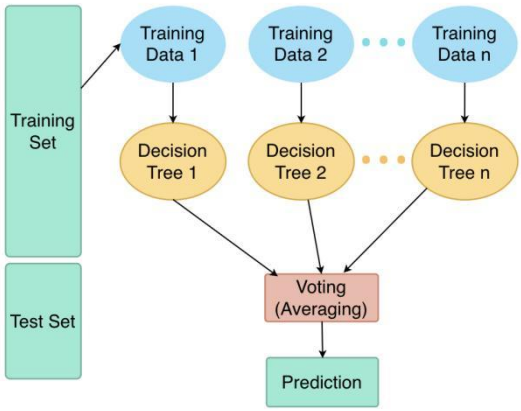
Şekil 2.1: Karar ağaçları algoritması çalışma prensibi

2.2.3 Rastgele Orman

Rastgele orman sınıflandırma algoritması, sınıflandırma ve regresyon problemlerinde kullanılabilen bir öğrenme yöntemidir.

Algoritma, birden fazla karar ağacı üreterek sınıflandırma işlemi esnasında sınıflandırma değerini yükseltmeyi hedefler. Rastgele orman algoritması birbirinden bağımsız olarak çalışan birçok karar ağacının bir araya gelerek aralarından en yüksek puan alan değerin seçilmesi işlemidir. Ağaç sayısı arttıkça kesin bir sonuç elde etme oranımız artmaktadır.

Karar ağaçları algoritması ile arasındaki temel fark, rastgele orman algoritmasında kök düğümü bulma ve düğümleri bölme işleminin rastgele olmasıdır.



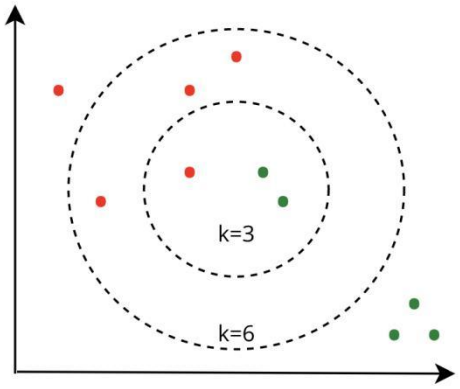
Şekil 2.2: Rastgele orman algoritması çalışma prensibi

2.2.4 K-En Yakın Komşu

KNN (K-Nearest Neighbors) tahmin edilecek değerin bağımsız değişkenlerinin oluşturduğu vektörün en yakın komşularının hangi sınıfta yoğun olduğu bilgisi üzerinden sınıfını tahmin etmeye dayanır.

Bilinmeyen bir örneklemin hangi sınıfa dahil olduğunu belirlemek için örüntü uzayını araştırarak bilinmeyen örnekleme en yakın olan k örneklemi bulur. Daha sonra bilinmeyen örneklem, k en yakın komşu içinden en çok benzediği sınıfa atanır. En yakın komşu bulunurken Euclidean, Manhattan, Chebyshev, Hamming, Minkowski, Mahalanobis, Haversiene, Levenshtein, Sørensen-Dice, Jacckard gibi uzaklık fonksiyonları kullanılmaktadır.

Sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacı ile K-en yakın komşu algoritması kullanılır [10].



Şekil 2.3: K-En yakın komşu algoritması çalışma prensibi

2.2.5 Multinominal Naive Bayes

Multinomial Naive Bayes algoritması, bayes teoremine dayanan çoğunlukla doğal dil işleme de (NLP) kullanılan olasılıksal bir makine öğrenmesi yöntemidir [16]. Bu algoritma, koşullu olasılıklara dayanarak hedef sınıftaki belirli bir değerin gerçekleşmesi ihtimalini inceler ve buna dayalı olarak hedef sınıfın değerini tahmin eder. Hedef sınıfı belirlerken en yüksek olasılığa sahip kararı seçmeyi amaçlar [17]. Örneğin e-posta veya gazete makalesi gibi bir metnin etiketini tahmin edilmesinde kullanılabilir.

$$P(c/x) = \frac{P(c/x)\,P(c)}{P(x)} \tag{2.2}$$

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

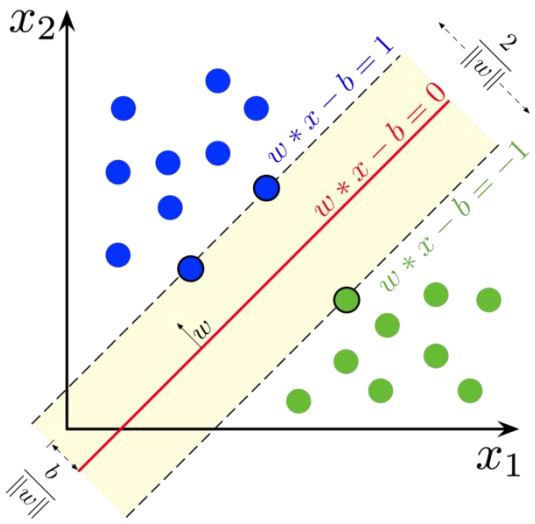
P(c): c olayının gerçekleşme olasılığı

P(x): x olayının gerçekleşme olasılığı

Algoritma sonuç olarak hedef sınıfın hangi değerinin gerçekleşme olasılığının ne olduğunu bildirir.

2.2.6 Destek Vektör Makineleri

Destek Vektör Makineleri (SVM), düzlem üzerindeki noktaların bir doğru veya hiper düzlem ile ayrıştırılması ve sınıflandırılmasıdır. SVM, bir düzlem üzerine yerleştirilmiş noktaları ayırmak için bir doğru çizer. Bu doğru, iki sınıfının noktaları için de maksimum uzaklıkta olmasını amaçlar. Bu doğrunun ±1'i arasında kalan bölgeye margin adı verilir. Margin ne kadar geniş ise iki veya daha fazla sınıf o kadar iyi ayrıştırılır.



Şekil 2.4: Destek vektör makinleri algoritması çalışma prensibi

w; ağırlık vektörü (θ1), x; girdi vektörü, b; sapmadır (θ0). Yeni bir değer için çıkan sonuç 0'dan küçükse, beyaz noktalara daha yakın olacaktır. Tam tersi, çıkan sonuç 0'a eşit veya büyükse, bu durumda siyah noktalara daha yakın olacaktır.

$$\breve{y} = \begin{cases} 0 \; if \; w^T . \; x + b < 0, \\ 1 \; if \; w^T . \; x + b >= 0 \end{cases} \tag{2.3}$$

2.3 Sınıflandırma Algoritmalarının Değerlendirilmesi

Model başarımını değerlendirirken kullanılan temel kavramlar hata oranı, kesinlik, duyarlılık ve F-ölçütüdür. Modelin başarısı, doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atılan örnek sayısı nicelikleriyle alakalıdır. Test sonucunda ulaşılan sonuçların başarım bilgileri karışıklık matrisi ile ifade edilebilir. Karışıklık matrisinde satırlar test kümesindeki örneklere ait gerçek sayıları, kolonlar ise modelin tahmin ettiği değerleri ifade eder.

		Öngörülen Sınıf	
		Sınıf = 1	Sınıf = 0
Doğru Sınıf	Sınıf = 1	a	b
	Sınıf = 0	c	d

a: TP (True Pozitif) c: FP (False Pozitif)
b: FN (False Negatif) d: TN (True Negatif)

Şekil 2.5: Karmaşıklık matrisi

2.3.1 Doğruluk

Model başarımının ölçülmesinde kullanılan en popüler ve basit yöntem, modele ait doğruluk oranıdır. Doğru sınıflandırılmış örnek sayısının (TP +TN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır.

$$\frac{TN + TP}{FN + TN + TP + FP} \tag{2.4}$$

2.3.2 Kesinlik

Kesinlik, sınıfı 1 olarak tahmin edilmiş True Pozitif örnek sayısının, sınıfı 1 olarak tahmin edilmiş tüm örnek sayısına oranıdır. Pozitif olarak tahmin ettiğimiz değerlerin gerçekten kaç adedinin pozitif olduğunu göstermektedir.

$$\frac{TP}{TP + FP} \tag{2.5}$$

2.3.3 Duyarlılık

Doğru sınıflandırılmış pozitif örneklem sayısının gerçek sınıfı pozitif olan tüm örneklemlerin sayısına oranıdır. Duyarlılık, gerçek pozitiflik oranı (True Positive Rate) olarak da adlandırılır.

$$\frac{TP}{TP + FN} \tag{2.6}$$

2.3.3 F-Ölçütü

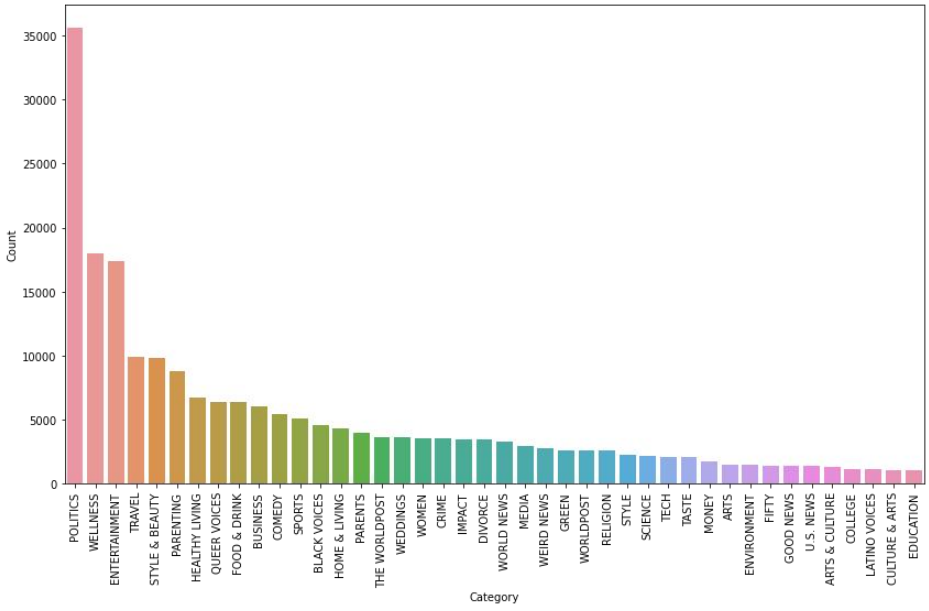
Kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verir. Bunun için f-ölçütü tanımlanmıştır. F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır.

$$\frac{2\,x\,Duyarlılık\,x\,Kesinlik}{Duyarlılık\,+\,Kesinlik} \tag{2.7}$$

Bölüm 3

Veri Seti

Veri seti makine öğrenmesi, veri analizi, yapay zeka, istatistik, veri görselleştirme, gibi bilimleri bir araya toplayan kaggle platformundan alınmıştır. Veri seti HuffPost'dan 2012-2022 tarihleri arasında toplanmış, 42 farklı kategoriden yaklaşık 210.000 haberden oluşmaktadır.



Şekil 3.1: Haberlerin kategorilere göre dağılımı

Veri setide ki her kayıt aşağıdaki öz niteliklerden oluşur.

- Kategori: Haberin ait olduğu kategori
- Başlık: Haberin başlığı
- Yazarlar: Haber'e katkıda bulunan yazarlar

- Adres: Haberin bağlantı adresi
- Özet: Haberin özeti
- Tarih: Haberin yayınlanma tarihi

	link	headline	category	short_description	authors	date
0	https://www.huffpost.com/entry/covid-boosters-...	Over 4 Million Americans Roll Up Sleeves For O...	U.S. NEWS	Health experts said it is too early to predict...	Carla K. Johnson, AP	2022-09-23
1	https://www.huffpost.com/entry/american-airlin...	American Airlines Flyer Charged, Banned For Li...	U.S. NEWS	He was subdued by passengers and crew when he ...	Mary Papenfuss	2022-09-23
2	https://www.huffpost.com/entry/funniest-tweets...	23 Of The Funniest Tweets About Cats And Dogs ...	COMEDY	"Until you have a dog you don't understand wha...	Elyse Wanshel	2022-09-23
3	https://www.huffpost.com/entry/funniest-parent...	The Funniest Tweets From Parents This Week (Se...	PARENTING	"Accidentally put grown-up toothpaste on my to...	Caroline Bologna	2022-09-23
4	https://www.huffpost.com/entry/amy-cooper-lose...	Woman Who Called Cops On Black Bird-Watcher Lo...	U.S. NEWS	Amy Cooper accused investment firm Franklin Te...	Nina Golgowski	2022-09-22

Şekil 3.2: Haber veri setinden örnek bir kesit

3.1 Veri Setinin İşlenmesi

Veri setinde kayıp gözlem olup olmadığı kontrol edilmiş kayıp gözlem olmadığı tespit edilmiştir. Veri setinde sınıf etiketi olarak kategori, öz nitelik olarak haber başlığı, özet ve yazarlar belirlenmiştir. Eğitim ve test veri setinde kullanılmayacak olan diğer öz nitelik değerleri silinmiştir.

Her öz nitelik sırasıyla küçük harfe çevrilip içerisinde geçen semboller ve özel karakterler temizlenmiştir. Natural Language Toolkit (NLTK) kütüphanesi kullanılarak öz niteliklere kelimelere ayırma (Tokenization), kelimede var olan ekleri kaldırıp morfolojik kök bulma(Lemmatization) işlemleri uygulanmıştır. Verilerin %75'i eğitim, %25'i ise test seti olarak ayrılmıştır.

	text	label
0	Over 4 Million Americans Roll Up Sleeves For O...	U.S. NEWS
1	American Airlines Flyer Charged, Banned For Li...	U.S. NEWS
2	23 Of The Funniest Tweets About Cats And Dogs ...	COMEDY
3	The Funniest Tweets From Parents This Week (Se...	PARENTING
4	Woman Who Called Cops On Black Bird-Watcher Lo...	U.S. NEWS
5	Cleaner Was Dead In Belk Bathroom For 4 Days B...	U.S. NEWS
6	Reporter Gets Adorable Surprise From Her Boyfr...	U.S. NEWS
7	Puerto Ricans Desperate For Water After Hurric...	WORLD NEWS
8	How A New Documentary Captures The Complexity ...	CULTURE & ARTS
9	Biden At UN To Call Russian War An Affront To ...	WORLD NEWS

	text	label
0	4 million american roll sleeve omicrontargeted...	U.S. NEWS
1	american airline flyer charge ban life punch f...	U.S. NEWS
2	23 funniest tweet cat dog week sept 1723until ...	COMEDY
3	funniest tweet parent week sept 1723accidental...	PARENTING
4	woman call cop black birdwatcher loses lawsuit...	U.S. NEWS
5	cleaner dead belk bathroom 4 day body found po...	U.S. NEWS
6	reporter get adorable surprise boyfriend live ...	U.S. NEWS
7	puerto ricans desperate water hurricane fionas...	WORLD NEWS
8	new documentary capture complexity child immig...	CULTURE & ARTS
9	biden un call russian war affront body charter...	WORLD NEWS

Şekil 3.3: Veri seti üzerinde yapılan ön işlem öncesi ve sonrası

3.2 Veri Setinin Sayısallaştırılması

Bu aşamada vektör uzay modelinden yararlanılmıştır. Vektör uzayları, metinlerin yapısal olmayan formdan sayısal hale getirilmesini sağlamaktadır. Bu yaklaşımda aslında her metnin vektör olarak temsil edildiği bir Kelime/Sözcük Çantası (BoW - Bag of Word) olup, her boyut ayrı bir terime karşılık gelmektedir. Her metin, mevcut kelimelerden oluşan MxN büyüklüğünde bir vektördür. Vektörler üst üste eklenerek döküman-terim matrisi oluşturulur. Bu matris, M adet haber ve n adet terimden oluşmaktadır. İlgili terimler haber içerisinde geçiyorsa o terimin ağırlık değeri sıfırdan farklı olur. Terimin ağırlık değeri, ilgili terimin metin üzerindeki etkisidir.

Terim ağırlığı hesaplanırken, terim sayma (Count Vectorizer) ve TF-IDF yaklaşımı kullanılmıştır. Count Vectorizer, ilgili terimin haber içerisinde ne kadar geçtini belirler. TF-IDF ise hem ilgili terimin haber içerisindeki sıklığına (Term Frequency – TF) hem de bütün haberler içerisindeki önemine bakar.

Bölüm 4

Modelin Eğitilmesi

Lojistik Regresyon, Karar Ağaçları, Rastgele Orman, K En Yakın Komşu, Multinominal Naive Bayes, ve Destek Vektör Makineleri sınıflandırma yöntemleri kullanılarak model eğitilmiştir. Sınıflandırma algoritmalarının performansları başarı metrikleri üzerinde karşılaştırılmıştır. Her bir model için aşağıda performans metrikleri verilmiştir.

Tablo 4.1: Algoritmaların sınıflandırma performanslarını gösteren tablo

	Metrikler			
	Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
Lojistik Regresyon	%69	%53	%69	%59
Karar Ağaçları	%53	%41	%43	%42
Multinominal Naive Bayes	%54	%26	%70	%29
Destek Vektör Makineleri	%70	%58	%64	%61
K- En Yakın Komşu	%12	%12	%67	%15
Rastgele Orman	%64	%47	%64	%52

Bölüm 5

Sonuç

Bu çalışmada haber veri seti üzerinde, Lojistik Regresyon, Karar Ağaçları, Rastgele Orman, K-En Yakın Komşu, Multinominal Naive Bayes ve Destek Vektör Makineleri algoritmaları kullanılmıştır.

Sonuçlar, makine öğrenimi algoritmalarının haber türünün belirlenmesindeki performansını analiz etmek amacıyla kullanıldığında farklı sonuçlar elde edildiğini göstermektedir. Lojistik regresyon algoritması, %69 doğruluk oranıyla diğer algoritmalar arasında en yüksek performansı gösterdi. Ayrıca, duyarlılık ve F-ölçütü değerlerinde de diğer algoritmalardan daha iyi sonuçlar elde etti. Bu, lojistik regresyonun haber türünün belirlenmesinde etkili bir seçenek olduğunu göstermektedir. Ancak, kesinlik değeri %53 olarak belirlendiğinden, daha yüksek bir kesinlik elde etmek için iyileştirmeler yapılabilir.

Karar ağaçları algoritması, %53 doğruluk oranıyla en düşük performans gösteren algoritma olarak belirlendi. Kesinlik, duyarlılık ve F-ölçütü değerleri de düşük seviyelerde bulunmaktadır. Bu sonuçlar, karar ağaçlarının haber türünün belirlenmesinde daha az etkili olduğunu göstermektedir. Modelin daha karmaşık yapıları ele alması ve daha fazla veri kullanması gerekebilir.

Multinominal naive Bayes algoritması, %54 doğruluk oranıyla diğer algoritmalar arasında ortalama bir performans göstermiştir. Ancak, kesinlik ve F-ölçütü değerleri oldukça düşüktür. Duyarlılık değeri %70 olarak belirlendi, bu da algoritmanın doğru pozitiflerin tespitinde daha iyi performans gösterdiğini göstermektedir. Bununla birlikte, düşük kesinlik ve F-ölçütü değerleri, algoritmanın doğru negatifleri tespit etmede zorluklar yaşadığını göstermektedir.

Destek vektör makineleri algoritması, %70 doğruluk oranıyla diğer algoritmalarla karşılaştırıldığında iyi bir performans göstermiştir. Ayrıca, kesinlik, duyarlılık ve F-ölçütü değerlerinde de tatmin edici sonuçlar elde etmiştir. Bu sonuçlar, destek vektör

makinelerinin haber türünün belirlenmesinde etkili bir yöntem olduğunu göstermektedir.

K-en yakın komşu algoritması, %12 doğruluk oranıyla en düşük performansı gösteren algoritmadır. Kesinlik ve F-ölçütü değerleri de düşüktür. Duyarlılık değeri %67 olarak belirlenmiştir, bu da algoritmanın doğru pozitifleri tespit etmede daha başarılı olduğunu göstermektedir. Ancak, düşük doğruluk oranı ve kesinlik değeri, algoritmanın genel olarak haber türünü doğru bir şekilde sınıflandırmada zorluklar yaşadığını göstermektedir.

Rastgele orman algoritması, %64 doğruluk oranıyla diğer algoritmalara kıyasla ortalama bir performans göstermiştir. Kesinlik ve F-ölçütü değerleri de ortalama seviyelerdedir. Duyarlılık değeri %64 olarak belirlendi, bu da algoritmanın doğru pozitifleri ve doğru negatifleri tespit etmede dengeli bir performans sergilediğini göstermektedir.

Lojistik Regresyon ve Destek Vektör Makineleri algoritmaları, diğerlerine kıyasla daha iyi sonuçlar elde etmiştir. Ancak, her bir algoritmanın avantajları ve dezavantajları göz önünde bulundurulmalı ve daha fazla veri ve model iyileştirmeleriyle sonuçların geliştirilmesi için çalışmalar yapılmalıdır.

Kaynaklar

[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] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[3] Chen, X., Zhang, Z., & Lian, Y. (2018). News Classification Based on Deep Learning. In 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 119–122). IEEE.

[4] Wang, Y., Huang, J., Jiang, S., & Wu, G. (2019). News Classification Method Based on Natural Language Processing and Machine Learning. In 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM) (pp. 391–395). IEEE.

[5] Smith, A. B., & Johnson, C. D. (2020). Support Vector Machines for News Type Classification. International Journal of Computer Science & Information Technology (IJCSIT), 12(2), 57–64.

[6] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3), 210–229. doi: https://doi.org/10.1147/rd.441.0206

[7] Mitchell, T. M. (1997). Machine Learning. In McGraw-Hill Science/Engineering/Math.doi:https://doi.org/10.1007/978-3-642- 21004- 4_10

[8] SAP. (2021). Was ist Maschinelles Lernen? https://www.sap.com/: https://www.sap.com/germany/insights/what-is-machine-learning.html

[9] Köse U (Nisan 2019) Yapay Zeka ve Geleceğin Siber Savaşları, Bilim ve Teknik, 76- 84.

[10] Kiani, F., Kutlugün, M. A., Çakır, M. Y., Yapay sinir ağları ve K-en yakın komşu algoritmalarının birlikte çalışma tekniği (ensemble) ile metin türü tanıma. İstanbul, 2017

[11] Nilsson, N. J., Introduction to machine learning an early draft of a proposed textbook. Stanford, 119 , 1998.

[12] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

[13] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). Applied Linear Statistical Models. McGraw-Hill Education.

[14] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. John Wiley & Sons.

[15] Şengür, D., Öğrencilerin akademik başarılarının veri madenciliği metotları ile tahmini. Fırat Üniversitesi, Eğitim Bilimleri Enstitüsü, Bilgisayar ve Öğretim Teknolojileri Eğitimi, Yüksek Lisans Tezi, 2013.

[16] S. Alqaraleh , "Efficient Turkish Text Classification Approach for Crisis Management Systems", Gazi University Journal of Science, 34(3), 718- 731, 2021, doi:10.35378/gujs.715296.

[17] Ö. Tonkal, H. Polat, “Traffic Classification and Comparative Analysis with Machine Learning Algorithms in Software Defined Networks”, Gazi University Journal of Science Part C: Design and Technology , 9 (1) , 71- 83 . doi: 10.29109/gujsc.869418.