Electronic Letters on Science & Engineering 18(2) (2022)
Available online at dergipark.org.tr/else

Comparison of Machine Learning Classification Algorithms: Example of Language Identification from Text

Furkan BAŞTÜRK[1], Hasan ŞAHİN[2*]

[1]Department of Computer Engineering, Bursa Teknik University, Bursa,Turkey
21435004021@btu.edu.tr, ORCID:0000-0002-3941-429X,
[2]Department of Industrial Engineering, Bursa Teknik University, Bursa,Turkey
h.sahin@btu.edu.tr, ORCID:0000-0002-8915-000X

Abstract: Artificial intelligence has accelerated its studies due to the rapid increase in technology and computing power today. In recent years, studies on natural language processing, whose importance is more prominent, have been boosted. The fact that people are in global communication has also increased this effect. There are many sub-branches of natural language processing. The beginning of all sub-branches takes place with language recognition. In studies, it is studied either through the determined language or by incorporating the language recognition process into the research. In this study, literature research and a sample study of language recognition have been done for the language recognition process. In addition, it aimed to contribute to the researchers who want to do natural language processing studies by offering the foresight to start their studies after performing the correct language recognition process.

Keywords: Machine Learning, Language Identification, Natural Language Processing, Naive Bayes

Makine Öğrenmesi Sınıflandırma Algoritmaların Karşılaştırması: Metinden Dil Tanıma Örneği

Özet: Günümüzde teknolojinin ve bilgisayar güçlerinin hızlı artışı sonucu yapay zekâ çalışmalarına da hız katmıştır. Son yıllarda önemi daha çok belli olan doğal dil işleme alanı üzerine çalışmalar artmaktadır. İnsanların global bir iletişim içinde olması da bu etkiyi arttırmıştır. Doğal dil işlemenin birçok alt dalı bulunmaktadır. Tüm alt dalların başlangıcı dil tanıma ile gerçekleşir. Çalışmalarda ya belirlenmiş dil üzerinden çalışılmakta ya da dil tanıma işlemini çalışma içerisine dahil ederek çalışılmaktadır. Bu çalışmada dil tanıma işlemi için literatür araştırması ve dil tanıma örnek bir çalışma yapılmıştır. Ayrıca doğal dil işleme çalışmaları yapmak isteyen araştırmacılara doğru dil tanıma işlemi yaptıktan sonra çalışmalarına başlaması öngörüsü sunularak katkıda bulunulması amaçlanmıştır.

Anahtar Kelimeler: Makine Öğrenmesi, Dil Tanıma, Doğal Dil İşleme, Naive Bayes

Reference to this paper should be made as follows (bu makaleye aşağıdaki şekilde atıfta bulunulmalı):
BAŞTÜRK, F. & ŞAHİN, H., 'Makine Öğrenmesi Sınıflandırma Algoritmaların Karşılaştırması: Metinden Dil Tanıma Örneği', Elec Lett Sci Eng , vol. 18(2), (2022), 68-78.

1. GİRİŞ

İnsanların var oluşundan bugüne kadar bir şekilde birbirleri ile iletişim kurmuşlardır. Bu iletişimi sesli, yazılı veya işaretler ile birbirleriyle iletişim kurarak gerçekleştirmişlerdir. Hatta hayvanlar ile bağ kuraraktan iletişim sağlamışlardır. Dünya da zamanla insan nüfusunun artmasıyla birlikte topluluklarda da artışlar olmuştur. Zamanla her toplum kendine yakın hissettiği veya kendine özgü belirlediği diller ile arasında iletişim kurmuştur. Bu dillerin oluşmasında aslında bulundukları çevresel ortamlar, komşu bulunan topluluklar vb. gibi etkenler diller üzerinde etkisi olmuştur. Karadeniz bölgesinin dağlık veya yamaç gibi arazi olmasından dolayı sürekli gidip gelinmesinin zor olduğu ve mesafeler arasında konuşmanın güç olmasından dolayı ıslık ile

* *Corresponding author; E-mail.:* h.sahin@btu.edu.tr

ISSN 1305-8614

iletişim kurma çözümünü bulmuşlardır. Islık kullanımını arttırmaları ile kendi aralarında yeni bir dil oluşturmuşlardır. Dünya da güncel olarak konuşulan 7000 dil olduğu bilinmektedir.

İletişimde en önemli etkenin ortak dil olduğu düşünülebilir. İnsanlar araştırmalarını, çalışmalarını ve buluşlarını bir şekilde insanlara duyurmak isterler ve çoğu zamanda bunu yazılı metinler, videolar vb. ile insanlığa duyururlar. Yayınlanan makalelere bakıldığında çoğunda kullanılan dil İngilizce olduğu görülmektedir. Birçok insana duyurmak istenilen bir konuyu İngilizce olarak yayınlamak gerekli diye düşünülebilir. Ama çalışmalarını kendi dili ile de anlatanlar bulunmaktadır. Çok önemli çalışmalar az bilinen diller arasında ise gözden kaçabilir. Bu sebeple dil tanımak çok önemlidir. Dünyada birçok farklı dili tanıyabilen ve konuşan insanlar bulunmaktadır. Bazı açıklamalara göre insan dil öğrenme sınırının bulunmadığından bahsetmektedir. Fakat bakıldığı zaman dünyada en çok dili konuşan insan olan Ziad Fazah 59 dil bilmektedir [1]. Dünyada konuşulan dillerin yanında işaret dilleri veya başka anlatım şekilleri düşünüldüğünde dünyada iletişim için kullanılan dillerin hepsini öğrenilmesi imkansızdır.

Aslında tüm dilleri öğrenmeden bizim yerimize öğrenip bize destek sağlayanlar bulunmaktadır. Bunlar çeviri yapan insan veya makine olabilirler. Bunlarda bir şekilde belirtilmiş olan diller arasında iletişim sağlamaktadır. Diller arasında çeviri yapabilmek için uygun dili bilen biri seçilmesi gerekmektedir. Ama tüm bu işleri gelişen teknoloji ile son zamanlarda önemi daha da artan yapay zekâ kavramı ile çözebiliriz. Yapay zekâ kendi başına öğrenebildiği ve bu süreci çok hızlı gerçekleştirdiği için yeterli çalışma ile tüm dilleri yapay zekaya öğretebilir. Bu sayede iletişim öğelerinin arasına yardımcı araç olarak yapay zekâ eklenmesi ile aynı dili bilmeyen kişilerin iletişim kurması sağlanabilir.

Bu çalışmada yapay zekâ kavramına genel bir bakış, yapay zekanın alt dallarından olan doğal dil işleme kavramına genel bir bakış, dil tanıma üzerine yapılan literatür araştırması ve dil tanıma için kullanılan makine öğrenmesi sınıflandırma algoritmaların karşılaştırılması için örnek bir uygulama yapılmıştır. Doğal dil işleme üzerine yapılacak çalışmalarda dil tanımanın önemi ve yapılacak çalışmaların başarılı dil tanıma sürecinden sonra olması gerektiğine dikkat çekmek amaçlanmıştır.

2. YAPAY ZEKA

2.1. Yapay Zekâ Nedir

'Yapay Zekâ' denilince işlevi açısından, insanların karar verme yeteneklerinin teknolojik bir biçimi anlaşılmaktadır. Ancak bu insanların davranışlarını kopyalamaya yönelik değildir. Yapay zekâ teknolojileri, endüstriyel süreçlerin verimliliğini ve etkinliğini artırmak için kullanılmak üzere tasarlanmıştır [22]. Yapay zekâ, belirli bir makinenin insan zekasını taklit edercesine çalışan bir yapıdır. Bu sistemde insana özgü bir durum olan düşünmeyi taklit etmeye çalışmaktadır. Bu çalışma sırasında deneyimlerinden yararlanarak anlam çıkarmada, genelleme yapmada ve sorunlara çözüm bulmada görev alır. Bir yandan da çalışmalarından kendine geri bildirim vermesi ile karar vermiş olduğu algoritmaları güncellemektedir [2]. Bu şekilde zekayı taklit ederek kendi geliştiren ve kendi zekâsı varmış gibi düşündüren makineler kurulmuş olur.

Yapay zekâ ne kadar yeni bir kavram olsa da günümüz teknolojisinin gelişmesiyle yapılan çalışmalarda da artış olmuştur. Çalışmalar hız kazanmış olsa da kontrolsüz gelişme çok sayıda hataya sebep olabilir. Yapay zekâ aslında adından belirli olabileceği gibi zekanın yani insanın bulunabileceği her alanda bulunabilir. İnsanın yapabileceği çoğu alanda bulunaktadır ve

gelecekte daha fazla alanda bulunacaktır. Bu durumun insanların işsizliği artıracağı konuşulsa da bazı mesleklerin kalkacağı ama yeni mesleklerin geleceği vurgulanmıştır [3].

2.2. Yapay Zekâ Çalışma Alanları

Yapay zekâ günümüzde birçok farklı alanda çalışmalar yapılmaktadır. Gelecekte daha çok alanda yapay zekayı görmemiz kaçınılmazdır. Şu anda yapılan çalışmalar ileride olacak çalışmalara öncülük etmekte ve yol göstermektedir. Günümüzde otomotiv, operasyon, sağlık finans, hukuk ve ses-görüntü alanları üzerinde çalışmalar vardır.

Sağlık sektöründe yapay zekâ kullanımının önceliği hastalık teşhisi üzerine olmuştur. Günümüz sağlık şartlarında çoğu hastalığın tedavisi kolaylaşmıştır ama en önemli etken doğru teşhis koyabilmektedir [4]. Ardında hastadan alınacak verilerin doğru alınması ve işlemesi önemli kriterlerdendir. Bu durumlar için de yapay zekâ kullanılarak birçok çalışma yapılmıştır.

Teknolojini gelişmesi endüstriyel sektörü de geliştirmiştir. Endüstri 4.0 gelişmesi ve yapay zekâ çalışmalarıyla otomotiv sektörü de olumlu anlamda etkileyerek araç üretilmesini de değişikliğe uğratmıştır. Bu değişiklik ile araçlara olan talebin tahmin edilmesi önem kazanmıştır. Ayrıca yapay zekayı araç içerisine aktarılması ile sürücüsüz araç çalışmaları yapılmaya başlanmıştır [5].

Yapay zekâ bankacılık sektöründe de önemli kazanımlar sağlamıştır. Bankacılık sektörünün en önemli yapısı müşteridir. Bu sebeple müşteri tarafında kazanımlar sağlayabilmek ve aynı zamanda sektörü hızlandırabilmek için kolay ve güvenli ödeme yöntemleri geliştirilmiştir. Mobil uygulamalar ile müşteri işlemleri kolaylaştırılmış ve herhangi bir insan desteği almadan birçok sorunu çözüm sağlayan sistem geliştirmişlerdir [6].

Hukuk alanında da yapay zekâ çalışmaları devam etmektedir. Yasa tahminlerini yapılmasında verilecek ceza süresini veya ücret miktarını tahmin edilmesi gibi farklı alanlar üzerinde çalışmalar yapılmaktadır [7].

Yapay zekâ araştırmaları arasında insanlığın en önemli düşlerinden biri hiç kuşku yok ki doğal diller arası çeviridir. Adalı'nın da (2012) [24] dediği gibi "düşlenen şey; bir kişinin ana dilinde konuşması, karşısındaki kişinin bunu kendi dilinde dinlemesidir". Uluslararası ve kültürlerarası bilgi aktarımın her geçen gün daha da önem kazandığı günümüz dünyasında bu düşü gerçekleştirmek artık bir zorunluluk haline gelmiştir [23].

3. DOĞAL DİL İŞLEME

3.1. Doğal Dil İşleme Nedir

Doğal dil işleme, yapay zekanın bir alt dalıdır. Gelişen teknoloji ile önemi hızla artmıştır. Makinelere doğal dillerin öğretilmesi amaçlanmıştır. Makinelerin insanların sesli veya yazılı iletişimlerini anlamasını ve anlam çıkarmasını sağlamak üzerine çalışılmaktadır. Hızla globalleşen dünyada insanları birbirinden ayıran önemli etken dillerin farklılığıdır. Bu farklılıkları en aza indirgemek ve insanların birbirleri ile hızlı iletişim kurması için doğal dil işleme çok büyük öneme sahiptir. Doğal dil işleme insanların kendi aralarında anlaşmak için kullandıkları dili insan-bilgisayar etkileşimini en üst düzeye çıkarabilmek veya farklı doğal dilleri kullanan insanlar arasında iletişimi güçlendirmek üzere çözümler üreten bilim alanıdır [21].

Doğal dil işlemede sesli veya yazılı olsun işlenecek içeriği metne çevrilir. Ardından bu metin kelimelerine ayrılarak dil bilimi işlemleri uygulanarak dil tespiti yapılır. Dil tespit edildikten sonra yapılacak çalışmalar dile özgü kurallar göz önüne alınır. Dil işleme sırasında birçok sorunla da karşılaşılır. Bu sorunlardan bazıları [8];

- Eklemeli diller de bulunan her ekten sonra kelimenin farklı anlama dönüşmesi,
- Kelimelerin cümle içindeki farklı yerlerde bulunmasından dolayı vurgunun farklılık göstermesi,
- Bir konuyu anlatan birden fazla kelime bulunduğu için makinelerin bu kelimelerden uygun olanı anmasıdır.

Bu gibi sorunların iyi çözüm getirebilmek için işlem yapılan dilin dil bilgisi yapısına hâkim olmak ona göre geliştirme yapılması gerekmektedir.

3.2. Doğal Dil İşleme Çalışma Alanları

Doğal dil işleme çalışmalarının ardından çok sayıda çalışmalar ortaya çıkmıştır. Bu çalışmalarla birlikte farklı alanlara odaklanan insanlar olmuştur. Metinlerden anlam çıkarılması, verilen konular ile metin oluşturulması, metin içeriklerinden duygu analizi yapılması, metin içerisindeki varlıkların tanınması ve otomatik çeviri sitemleri gibi birçok konu üzerinde çalışmalar yapılmıştır [9]. Gelecekte de daha çok çalışma olması kaçınılmazdır. İleride çok farklı konularda başarılı çalışmalar yapılacağı beklenmektedir.

Çalışmaların önemli gördüğümüz doğal dil işlemenin alt dallarından olan dil tanıma üzerine de birçok çalışma gerçekleştirilmiştir.

4. YAPILAN ÇALIŞMALAR

Doğal dil işlemenin alt dalları olan dil tanıma çalışmaları üzerine yapılan literatür araştırması yapılmıştır.

Tarcan ve Çakar (2008) dil bilimsel yöntemleri açıklamışlar. Dil tanıma üzerine çalışmalarını yapmışlardır. Bu çalışmada N-gram metodunu kullanarak web sitesinin dili tanıma uygulaması geliştirmişlerdir. Bu çalışma sonucu ile URL bilgisi girilen web sitesinin Türkçe olup olmadığına karar veren uygulamayı gerçekleştirmişlerdir [10].

Yavanoğlu ve Sağıroğlu (2010) istatiksel dil tanıma metotlarının yanında farklı bir bakış açısı kazandırmayı amaçlamışlardır. Çalışmalarında 15 dil üzerinde çalışmışlardır ve yapay sinir ağlarını kullanmışlardır. Bunun yanında birleşim tespit yöntemi adını verdiklerini yeni yöntem geliştirmişlerdir. Bu yöntemde algoritmanın ezberlemesini azaltmak amacıyla dillerdeki alfabelerin en az 2 dilde bulunma kuralını belirlemişlerdir. Bu çalışma sonucunda Word ve html içerisindeki metinlerin dil tanıma başarısını %99 olduğunu belirtmişlerdir [11].

Yavanoğlu ve Sağıroğlu (2012) çalışmalarında önceki çalışmalarıyla beraber çalışacak yeni yöntem geliştirerek dil tanıma doğruluk oranını artırmak istemişlerdir. Çalışmalarında 15 dili tanımak ve ardından da 64 dile otomatik çevirisinin yapılması amaçlanmıştır. Çalışmalarında yapay sinir ağları, birleşim tespit yöntemi ve yeni geliştirdikleri kesişim tespit yöntemini kullanmışlardır. Kesişim tespit yöntemi, klasik Latin alfabesinde ve modern Latin alfabesinde yer alan ve en az 5 alfabe ortak olan harflerin kullanılması kuralını belirlemişlerdir. Bu çalışma

sonucunda başarı oranını %100 olduğunu belirtmişlerdir. Ama çalışmanın ortalama sonuçları Word için %52-%98, PDF için %52-%99 ve HTML için %75-%99 dur [12].

Kaya ve Ertuğrul (2016) dil tanıma işlemleri için görüntü işleme metotlarından olan yerel ikili örüntüler metodunu kullanmışlardır. Veri ön işlemenin ardından UTF-8 formatına kullanılmış ardından 1 boyutlu YİD kullanmışlardır. Sonuçların karşılaştırılması için yapay sinir ağları, destek vektör makinesi ve çok terimli Naive Bayes yöntemi kullanmışlardır. Çalışmanın sonucunda α ve β parametrelerinde anlamlı ve sürekli ilişki tespit edememişlerdir. 4 farklı veri setinde %86,20, %92,75, %100 ve %89,77 sonuçları elde etmişlerdir. Cümlenin uzunluğunun artması başarı oranını arttırdığını savunmuşlardır [13].

Aslanyürek ve Mesut (2021) dil tanıma işleminde farklı öznitelik seçme yöntemleri dil tanıma algoritmaları karşılaştırılmışlardır. 4 farklı veri seti kullanmışlar. Veri önişleme ardında dilbilimsel yöntemler uygulanarak algoritmaların dil tespitinin sonuçları karşılaştırılmıştır. Çalışmanın sonucunda 0.2 KB ve daha küçük metinlerde Naive Bayes algoritmaları ile SelectFromModel – Lojistik regresyon uygulandığı zaman diğerlerine göre daha yüksek doğruluk değeri vermektedir. 0.2 KB ve 0.5 KB arasındaki metinlerin tamamında SelectFromModel – Lojistik regresyon yöntemi yüksek doğruluk değeri vermiştir. Ayrıca en yüksek sınıflandırma algoritması Fasttext olduğunu belirtmişlerdir [14].

Çelik ve Odabaş (2020) işaret dilini metine dönüştürülmesi çalışılmıştır. El hareketlerini görüntü işleme ile algılanması sonrasında Konvolüsyonel Yapay Ağlar metodu ile yaptıkları çalışmada sabit hareketlerin tahin edilmesin başarılı olduğu ancak hareketli hareketlerin tahmin edilmesinin başarı oranı %53 olduğunu belirtmişlerdir. Bu sorunu da Uzun Kısa Süreli Bellek metodu ile çözmüşlerdir. CNN ile LSTM kullanılması ile başarı oranını %97 olduğunu belirtmişlerdir [15].

Zhu ve ark., (2008) el ile yazılmış veya baskı alınmış şekilde bulunan metinlerinin dil tanınması amaçlamışlardır. Bu çalışma için Canny edge detector yöntemini kullanmışlardır. Elde ettikleri her bitiş 3 segmenti TAS özelliği olarak belirtmişlerdir. Bu özelliklere kümeleme ve maksimize uygulamışlardır. Çalışmalarının sonucun da Çince ve Japonca arasında ayrım yapmada zorlandığını belirtmişlerdir. Ayrıca doku tabanlı LBP farklı veri kümesindeki farklılıkları tanıyamamıştır. Sonuç olarak 8 farklı dil için dil tanıma başarısı %95,6 olarak belirtilmiştir [16].

Bruno Martins ve MÆrio J. Silva (2005) web sayfasında bulunan metinlerin diğer metinlere göre dil tespitinin daha zor olduğunu belirtmişlerdir. Bu sebeple Web sayfasının kategorizasyonu için sezgisel yöntem geliştirilerek web sayfasının dilini tahmin etmeyi amaçlamışlardır. Sayfalarda bulunan meta verilerinin hataya sebep olduğunu belirtmişlerdir. Çalışmalarında N-gram metodunu kullanmışlardır. Başlıklar metnin belirten kısım oldukları için ana dili vurguladığını düşündükleri için başlıkları 3 kez, açıklama içeriklerini 2 kez saymışlardır. 40 karakterden az olan metinleri tanımsız olarak tanımlamışlardır. Ayrıca sayfa içerisinde birden çok dil bulunduğunda en uzun metnin ağırlığını arttırarak N-gram metodunu tekrardan çalıştırırlar. Çalışmanın sonucunda Kuzey Avrupa dillerini tanımasını genel olarak karıştırmıştır. Ama Portekiz dilini %92 başarı oranında tanımlamıştır [17].

Rafael Dueire Lins ve Paulo Gonçalves Jr (2004) kapalı dil bilgisine sahip yazılı metinlerin dil tanıması üzerine yeni algoritma geliştirmeyi amaçlamışlardır. Çalışmalarını Portekizce, İspanyolca, Fransızca ve İngilizce dilleri üzerine yapmışlardır. Geliştirdikleri algoritma içerisinde tüm sözlükte bulunan kelimelerin toplamının her sözcüktü bulunan kelimelere oranı LimID olarak tanımlamışlardır. Girdi metin içinde aynı formül uygulanarak LimPAL tanımlamışlardır. Bu değerleri her dil için oranlamasını uygulamışlardır. Formül sonucuna göre

İspanyolca, Fransızca, Portekizce ve İngilizce sıralamasına göre dili kabul eder. Çalışma sonucunda HTML metinleri için %80 başarı oranı elde etmişlerdir. Metin içeriği belirtilen dillerden ise dil tanıma süresi 0,03sn sürmekte ve metin belirtilen dillerden değil ise 0,2sn tahmin etmektedir. Yazılı metinler için ise %90 başarı oranı elde etmişlerdir. Dil tanıma süresi belirtilen dillerden ise 0,01sn'de dil tanımakta eğer farklı dil ise 0,2sn'de tanıma yapabilmektedir [18].

Mohammad. M. AlyanNezhadi, Majid Forghani ve Hamid Hassanpour (2017) sinyal işleme teknikleri kullanılarak dil tanıma işlemine yeni yaklaşım getirmeyi amaçlamışlardır. Ön işleme aşamasında ardışık kelimeler arasına boşluk ekleyerek sinyal işlemenin çözünürlüğünü değiştirmişlerdir. Bu değişiklik ile sinyal üzerinde daha kararlı sonuçlar elde edilmektedir. Metni UFT-8 formatına çevirdikten sonra dalgacık dönüşümü uygulamışlardır. Bu işlemden sonra 32 özellik belirlenmiştir. Belirledikleri 32 özelliği yapay sinir ağlarına giriş olarak belirtmişlerdir. Çıktı olarak seçmiş oldukları 7 dili belirlemişlerdir. Sonuçların doğrulaması için 10 kat çapraz doğrulama uygulamışlardır. Çalışmanın sonucunda metne 1 boşluk eklendiği zaman %91,2 doğruluk oranı, metne 5 boşluk eklendiği zaman %97,1 doğruluk oranı, metne 7 boşluk eklendiği zaman %97,4 doğruluk oranı, metne 12 boşluk eklendiği zaman doğruluk oranı %96,3 sonucuna varmışlardır. Metne 10'dan fazla boşluk eklendiği zaman doğruluk oranın düştüğünü belirtmişlerdir [19].

Dil tanıma alanında gerçekleştirilen yapay zekâ makaleleri incelendikten sonra çalışmanın amacı, kullanılan yöntem, literatüre katkısı, elde edilen sonuç ve çalışmanın önerileri bilgileri ile Tablo 1'de aktarılmıştır.

Tablo 1. İncelenen çalışmalara ilişkin özet bilgiler

Yazar(lar) Adı	Yıl	Makale Adı	Çalışmanın Amacı	Kullanılan Yöntem	Literatüre Katkısı	Elde Edilen Sonuç
Ahmet TARCAN Fahri ÇAKAR	2008	Bilgisayarli Dil Tanimlam da Dilbilimsel Yaklaşimlar ve Bir Yazilim Denemesi	Ana dili Türkçe olmayan kullanıcılar için web sitesinin dil tanımlanması yapılmıştır.	N – Gram, Delphi	Dilbilimsel ölçütler algoritmaya eklendiğinde dilleri daha kolay ayırt ettiklerini belirtmişlerdir	Uygulamaya URL girilerek web sitesinin Türkçe olup olmadığına karar vermektedir
Uraz YAVANOĞLU Şeref SAĞIROĞLU	2010	Web Tabanli Otomatik Dil Tanima ve Çevirme Sistemi	Web ortamındaki içerikleri otomatik olarak dil tanımlanması yapılmasını ardında istenilen dile otomatik çevirmesini çalışmışlardır.	Yapay sinir ağları, Birleşim Tespit Yöntemi, Celbi Microsoft Word Otomasyon kütüphanesi	Yeni dil tanıma yöntemi geliştirmişlerdir	HTML ve Word dosyasının içeriğini 15 farklı dilde tahmin edebilme başarısı %99 seviyesinde bulunmuştur
Uraz YAVANOĞLU Şeref SAĞIROĞLU	2012	Zeki Doküman Dili Sınıflandırma ve Web Tabanlı Çeviri Sistemi	WORD, PDF ve HTML üzerinden 15 farklı dilin tanımlanması ardından 64 farklı dile otomatik çevirisi amaçlanmıştır.	Yapay sinir ağları, Birleşim Tespit Yöntemi, Kesişim Tespit Yöntemi	Yeni dil tanıma yöntemi geliştirmişlerdir. Önceki çalışmalarını birleştirerek doğruluk oranı artırmışlardır.	Yeni sitemin doğruluk oranı %100 dür. Dil tanıma başarı ortalamaları ise Word için %52-%98, PDF için %52-%99 ve HTML için %75-%99 dur
Yılmaz KAYA, Ömer Faruk ERTUĞRUL	2016	Doküman Dili Tanıma için Yeni Bir Öznitelik Çıkarım Yaklaşımı: İkili Desenler	Dil tanıma işlemleri için ikili desenler kullanılarak yeni bir dil tanıma yaklaşımı önermişlerdir.	UFT-8 Formatı, Yerel ikili örüntüler, Yapay sinir ağları, Destek vektör makineleri, Çok terimli Naive Bayes	Dil tanıma için tümüyle farklı ve yeni yaklaşımı başarılı şekilde önerilmiştir.	4 farklı veri setinde %86,20, %92,75, %100 ve %89,77 sonuçları elde etmişlerdir. Cümlenin uzunluğunun artması başarı oranını arttırdığını

						savunmuşlardır
Murat ASLANYÜREK Altan MESUT	2021	Kısa Metinleri Yazıldıkları Dile Göre Sınıflandırma ve Farklı Öznitelik Seçim Yöntemlerinin Uygulanması	Dil tanıma işleminde farklı öznitelik seçme yöntemlerini kullanarak dil tanıma yöntemlerinin karşılaştırılmasını amaçlamışlardır	Naive Bayes, Karar Ağaçları, K-En Yakın Komşu, Destek Vektör Makinesi, Langdetect, Fasttext, Terim Frekansı Ters Doküman Frekansı, Ki-Kare, SelectFromModel Lojistik Regresyon	Dilbilimsel ölçütler algoritmaya eklendiğinde dilleri daha kolay ayırt ettiklerini belirtmişlerdir	4 Veri setinde yapılan çalışma sonucunda 0,2 KB ve daha küçük metinlerde M-NB ve B NB ile SelectFromModel-LR daha yüksek doğruluk verirken 0,2 KB-0,5 KB arası metinlerde tamamında SelectFromModel-LR daha yüksek doğruluk değeri vermiştir. En yüksek doğruluk değerine sahip sınıflandırma yöntemi Fasttext sonucuna varılmıştır.
Özer ÇELİK, Alper ODABAŞ	2020	Sign2Text: Konvolüsyonel Sinir Ağları Kullanarak Türk İşaret Dili Tanıma	Herhangi bir sensör kullanılmadan işaret dilini metne dönüştürülmesi amaçlanmıştır.	Konvolüsyonel Yapay Ağlar, Uzun Kısa Süreli Bellek, Görüntü işleme metotları	İşaret dilini metne dönüştürülmesi için CNN + LSTM kullanarak yeni model geliştirmişlerdir	CNN + LSTM modelinde %97 başarı elde edilmiştir
Guangyu ZHU, Xiaodong YU, Yi Lİ, David DOERMANN	2008	Unconstrained Language Identification Using A Shape Codebook	El yazısı ve makinede basılmış metinlerden dil tanımasına yeni bir yaklaşım önermeyi amaçlamışlardır.	Canny Edge Detector, LBP, SVM	Gerçek durumda bulunan el yazısı ve makine baskısı metinlerin dil tanıma işlemine yeni hesaplamalar katmışlardır.	Geliştirdikleri yeni yaklaşım ile 8 dil için dil tanıma başarı oranları %95,6 olmuştur
Bruno MARTİNS, MÆrio J. SİLVA	2005	Language Identification in Web Pages	Web sayfasının kategorizasyonu için sezgisel yöntem geliştirilerek web sayfasının dilini tahmin etmeyi amaçlamışlardır.	N – Gram,	Web sayfasından N-gram yöntemini kullanarak dil tanıma gerçekleştirmiştir.	Kuzey Avrupa dilleri genel olarak karıştırmıştır ve kötü sonuçlar elde etmiştir. Ama Portekiz dil tespitinde %92 başarı sağlamıştır
Rafael Dueire LİNS, Paulo GONÇALVES	2004	Automatic language identification of written texts	Kapalı dil bilgisine sahip olan yazılı metinlerin dil tanımasında yeni algoritma önerilmesi amaçlanmıştır	Yeni algoritma geliştirmişlerdir.	Portekizce, İspanyolca, Fransızca ve İngilizce dilleri için HTML ve yazılı metinlerde dil tanıma için başarılı yeni metot önerilmiştir	HTML sayfalarında başarı oranı %80, yazılı metinlerde başarı oranı %90'dır. HTML sayfaları belirtilen dillerde ise dil tanıma süresi 0,03sn, yazılı metinlerde ise 0,01sn sürdüğünü belirtmişlerdir.
Mohammad. M. ALYANNEZHADİ, Majid FORGHANİ, Hamid HASSANPOUR	2017	Text Language Identification Using Signal Processing Techniques	Sinyal işleme teknikleri kullanılarak dil tanıma işlemine yeni yaklaşım önerilmesi amaçlanmıştır.	Dalgacık dönüşümü, Yapay sinir ağları, Çapraz doğrulama	Metinlerin alt bant katsayılarına sinyal işleme uygulanması ile dil tanıma işlemine yeni yöntem önerilmiştir	Metne 1 boşluk eklendiğinde %91,2, 5 boşluk eklendiğinde %97,1, 7 boşluk eklendiğinde %97,4 ve 12 boşluk eklendiği zaman %96,3 doğruluk sonucuna varmışlardır

5. UYGULAMA

Uygulama çalışmasında beş adım takip edilmiştir. İlk aşamada veri seti oluşturulması, ikinci aşamada metin işleme için gerekli olan veri ön işleme işlemleri, üçüncü aşamada makine öğrenmesi sınıflandırma algoritmaları kullanılarak model oluşturulması, dördüncü aşamada algoritma sonuçlarının karşılaştırılması, son aşamada modellere veri seti dışından veri verilmesi ile modellerin tepkileri çalışılmıştır.

5.1. Veri Seti

Veri seti için açık kaynak veri seti kullanılmıştır. Bunun için Kaggle web sitesi üzerinde bulunan veri seti kullanılmıştır. Bu veri seti içerisinde 22 dil bulunmaktadır. Her dil için 1000 satır toplamda 22.000 satır bulunmaktadır [20].

5.2. Veri Ön İşleme

Veri seti içerisindeki benzersiz satırlar kontrol edildiğinde 21.859 satır bulunmaktadır. Algoritmaların ezberleme yapmaması için kopya olan satırlar veri setinden çıkarılır.

Metin işleme sırasındaki ön işleme işlemleri için alfanümerik olmayan karakterler çıkartılır. Metin kelime listesine dönüştürülür. Kelimeler üzerinde köklendirme denilen kelimenin kök haline çevrilmesi sağlanır. Ardından tüm kelimeler küçük harfe dönüştürülür.

5.3. Modelleme

Ön işleme yapılan veri setinin %80 eğitim, %20 test verisi olarak kullanılmak üzere rastgele paylaştırılmıştır. Modelleme işleminde makine öğrenmesi sınıflandırma algoritmalarından Bernoulli Naive Bayes, Multinomial Naive Bayes, Decision Tree, Support Vector Machine, K-Nearest Neighbors ve Random Forest algoritmaları kullanılmıştır. K-Nearest Neighbors algoritması için N katsayısı 1000 olarak belirlenmiştir.

5.4. Sonuç Karşılaştırılması

Oluşturulan modellerin eğitim sonrasında test verisi kullanılarak algoritmaların dil tahmin etmesindeki doğruluk sonuçları ve eğitim ile test işlemleri sırasındaki işlem süresi bilgileri çıktı olarak tutulmaktadır. Ardından doğruluk değeri önemli bir kriter olduğu için ağırlık katsayısı 2, işlem süresi ağırlık katsayısı 1 olarak belirlenmiştir. Bu katsayı oranları ile algoritmaların birbirleri arasındaki başarı sıralaması oluşturulmuştur. Tablo 2'de algoritmaların başarı sıralamasına göre doğruluk oranları ve işlem süreleri verilmiştir.

Tablo 2. Algoritmaların doğruluk ve işlem süresine göre iyiden kötüye sıralanması

Başarı Sırası	Algoritma Adı	Doğruluk Oranı	İşlem Süresi(sn)
1	Multinomial Naive Bayes	%95.151	0.14
2	Random Forest	%91.857	109.46
3	Support Vector Machine	%90.256	46.01
4	Bernoulli Naive Bayes	%88.632	0.24
5	Decision Tree	%89.021	13.46
6	K - Nearest Neighbors	%21.477	1.61

5.5. Modellerin Uygulanabilirliği

Modellerin uygulanabilirliği için tüm modellere veri seti dışından kısa cümle ve paragraf bilgisi verilerek modellerin sonuçları karşılaştırılmıştır. Dil tahmin çıktılarına bakıldığında kısa cümle verilmesinde doğru tahmin oranının oldukça kötü olduğu belirlenmiştir. Paragraf verildiği zaman ise doğru tahmin oranının yüksek olduğu görülmüştür. Tablo 3'te sonradan modele verilen paragrafların bilgisi yer almaktadır. Paragraflar güncel haber içeriklerinden alınmıştır.

Tablo 3. Veri seti dışında kullanılan paragraf içerikleri

Sıra No	Paragraf
1	Taburenin kendisi için manevi değere sahip olduğunu söyleyen sahibi ise geri getirilmesini istedi.
2	Eynesilin çayıyla meşhur olduğuna dikkati çeken Tufanoğlu, Burası çay memleketi, 5 çay fabrikası var. Çayımızı tanıtmak, temsil etmek istedik. Semaverde çay içmenin de ayrı bir manası var, insanlar semaverin etrafında bir araya geliyor. Bu açıdan da semaver iyi oldu. ifadelerini kullandı. Ayhan Tufanoğlu, caminin şadırvan gibi bazı bölümlerindeki çalışmaların ise devam ettiğini sözlerine ekledi.

Tablo 4'te verilen paragrafların modellere göre yapılan dil tahminleri yer almaktadır.

Tablo 4. Veri seti dışında verilen paragrafların dil tahmini

Algoritma Adı	1.Paragraf Dil Tahmini	2.Paragraf Dil Tahmini
Multinomial Naive Bayes	Turkish	Turkish
Random Forest	Turkish	Turkish
Support Vector Machine	Japanese	Turkish
Bernoulli Naive Bayes	Japanese	Turkish
Decision Tree	Turkish	Turkish
K - Nearest Neighbors	Japanese	Japanese

6. SONUÇ

Ülkemizde ve dünyada doğal dil işleme konusu giderek popüler hale gelmektedir. Özellikle akıllı telefonların da hayatımıza girmesiyle, kişisel asistanlık yapan, insanlardan aldığı sesli komutları yerine getiren uygulamalar bu konuya ilgiyi artırmaktadır. Bu çalışmada, dil tanımada yapay zekâ kavramı ve bu alanda yapılmış çalışmaların literatür araştırması yapılmıştır. Son yıllarda popüler olan doğal dil işleme kavram üzerine yeni çalışmalar bulunsa da dil tanıma çalışmaları eski yıllara oranla azdır. Yapılacak çalışmaların çoğu için doğru ve etkili çalışan dil tanıma alt yapısı oluşturulmalıdır.

Çalışmaları incelediğimizde etkili modellerin makine öğrenmesi algoritmaları ve yapay sinir ağları tercih edilmektedir. Ayrıca dil tanıma işlemi için yaptığımız çalışmamızda yapılan eğitim ve test işlemleri sonucunda göre en iyi sonucu %95.151 başarı oranı ile Multinomial Naive Bayes, en kötü sonucu %21.477 başarı oranı ile K - Nearest Neighbors algoritması sağlamıştır. Algoritmalar arasından en hızlı tahmin eden 0.14sn ile Multinomial Naive Bayes, en yavaş tahmini eden 109.46sn ile Random Forest algoritması olmuştur. Belirlenen ağırlık katsayısı ile değerlendirildiğine en iyi sonucu Multinomial Naive Bayes, en kötü sonucu K - Nearest Neighbors algoritması sağlamıştır.

Model çalışması tamamlandıktan sonra veri setinde bulunmayan veriler ile model kontrol edildiğinde algoritmaların genelinde kısa paragrafta dil tahmin edilmesi zor olduğu gözlemlenmiştir. Sonuç olarak yapay zekanın ve doğal dil işlemenin gelecek yıllarda önemli bir konumda olacağı ön görülmektedir.

Kaynakça

[1] "Ziad Fazah - Wikipedia," Jun. 20, 2022. https://en.wikipedia.org/wiki/Ziad_Fazah (accessed Jun. 20, 2022).

[2] V. V. Nabiyev, Yapay Zeka: İnsan Bilgisayar Etkileşimi. Ankara: Seçkin Yayıncılık, 2010. Accessed: Jun. 20, 2022.

[3] Ö. ÇARK, "Dijital Dönüşümün İşgücü Ve Meslekler Üzerindeki Etkileri," International Journal Entrepreneurship and Management Inquiries, vol. 4, no.1, pp. 19–34, 2020.

[4] B. Akalın and Ü. Veranyurt, "Sağlık Hizmetleri ve Yönetiminde Yapay Zekâ ," Acta Infologıca, vol. 5, no. 1, pp. 231–240, 2021, doi: 10.26650/acin.850857.

[5] Y. E. Sürmen and E. Güler, "Endüstri 4.0 Ve Otomotiv Endüstrisi: Bursa İli Swot Analizi İle Değerlendirilmesi" Yayınlanmamış Yüksek Lisans Tezi, Bursa Uludağ Üniversitesi Fen Bilimleri Enstitüsü, 2019, Bursa.

[6] E. LastNameGümüş, B. Medetoğlu, and S. Tutar, "Finans ve Bankacılık Sisteminde Yapay Zekâ Kullanımı: Kullanıcılar Üzerine Bir Uygulama," Bucak İşletme Fakültesi Dergisi, vol. 3, no. 1, 2020.

[7] W. Yang, W. Jia, X. Zhou, and Y. Luo, "Legal Judgment Prediction via Multi-Perspective Bi-Feedback Network," Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019.

[8] S. E. Seker, "Doğal Dil İşleme(Natural Language Processing)," YBS Ansiklopedi, vol. 2, no. 4, 2015, Accessed: Jun. 20, 2022.

[9] D. Küçük and N. Arici, "Doğal Dil İşlemede Derin Öğrenme Uygulamaları Üzerine Bir Literatür Çalışması," Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi, vol. 2, no. 2, pp. 76–86, 2018.

[10] A. Tarcan And F. Çakar, "Bilgisayarli Dil Tanimlamada Dilbilimsel Yaklaşimlar ve Bir Yazilim Denemesi," Elektronik Sosyal Bilimler Dergisi, vol. 7, no. 26, pp. 64–70, 2008, Accessed: Jun. 20, 2022.

[11] U. Yavanoğlu And Ş. Sağıroğlu, "Web Tabanlı Otomatik Dil Tanıma Ve Çevirme Sistemi," Gazi Üniv. Müh. Mim. Fak. Der., vol. 25, no. 3, pp. 483–494, 2010, Accessed: Jun. 20, 2022.

[12] U. Yavanoğlu and Ş. Sağıroğlu, "Zeki doküman dili sınıflandırma ve web tabanlı çeviri sistemi," Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi, vol. 28, no. 4, pp. 329–342, 2012.

[13] Y. Kaya and Ö. F. Ertugrul, "Doküman dili tanıma için yeni bir öznitelik çıkarım yaklaşımı: İkili Desenler," Journal of the Faculty of Engineering and Architecture of Gazi University, vol. 31, no. 4, pp. 1085–1094, 2016, doi: 10.17341/GAZIMMFD.278463.

[14] M. Aslanyürek and A. Mesut, "Kısa Metinleri Yazıldıkları Dile Göre Sınıflandırma ve Farklı Öznitelik Seçim Yöntemlerinin Uygulanması," Journal of Investigations on Engineering & Technology, vol. 4, no. 2, pp. 36–46, 2021.

[15] Ö. Çelik and A. LastNameOdabaş, "Sign2Text: Konvolüsyonel Sinir Ağları Kullanarak Türk İşaret Dili Tanıma," Avrupa Bilim ve Teknoloji Dergisi, no. 19, pp. 923–934, 2020, doi: 10.31590/ejosat.747231.

[16] G. Zhu, X. Yu, Y. Li, and D. Doermann, "Unconstrained Language Identification Using A Shape Codebook," Computer Science, 2008, Accessed: Jun. 20, 2022.

[17] B. Martins and M. J. Silva, "Language identification in Web pages," Proceedings of the ACM Symposium on Applied Computing, vol. 1, pp. 764–768, 2005, doi: 10.1145/1066677.1066852.

[18] R. D. Lins and P. Gonçalves, "Automatic language identification of written texts," Proceedings of the ACM Symposium on Applied Computing, vol. 2, pp. 1128–1133, 2004, doi: 10.1145/967900.968129.

[19] M. M. A. Nezhadi, M. Forghani, and H. Hassanpour, "Text language identification using signal processing techniques," Proceedings - 3rd Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS 2017, vol. 2017-December, pp. 147–151, Mar. 2018, doi: 10.1109/ICSPIS.2017.8311606.

[20] "Language Identification dataset Kaggle." https://www.kaggle.com/datasets/zarajamshaid/language-identification-datasst (accessed Jun. 20, 2022).

[21] https://yapayzeka.itu.edu.tr/arastirma/dogal-dil-isleme Erişim Tarihi: 10.10.2022

[22] Şahin, H., Sosyal Hizmet Çalışmaları- 2 İnsan Değeri Ve Değerlemeleri / Yapay Zekâ ve İnsan Değeri, EKİN Basım Yayın Dağıtım Osmangazi / BURSA, 2021.

[23] Kuşçu, E. (2015). Çeviride Yapay Zeka Uygulamalari. Atatürk Üniversitesi Kazım Karabekir Eğitim Fakültesi Dergisi, (30), 45-58.

[24] Adalı, E. (2016). Doğal Dil İşleme. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5(2). Retrieved from https://dergipark.org.tr/en/pub/tbbmd/issue/22245/238797