Makine öğrenmesi yöntemleri ile türk lehçelerinin tahminlenmesi

Yasin ŞAHİN[1]

[1]*Kocaeli Üniversitesi*
Bilişim Sistemleri Mühendisliği
[1]181307026@kocaeli.edu.tr

I. INTRODUCTION

Günümüzde makine öğrenmesi algoritmaları sayesinde çoğu sorunun çözümü için kullanılmaktadır. Bu araştırma projesinde türk lehcelerinin sınıflandırılması için 5 farklı makine öğrenmesi ve derin öğrenme yöntemi kullanılmıştır. Yöntemler arasında en iyi sonuc evrişimsel sinir ağı yönteminde gerçekleştirildi. Doğruluk değerleri evrişimsel sinir ağı %97.5, destek vektör makinesi %88, yapay sinir ağı %95.4, transfer öğrenmesi %85.4, Naive bayes %76.7 değerlerinde sonuçlanmıştır.

anahtar kelimeler **lehce tahmini, ses bölümlemei**

II. LITERATÜR TARAMASI

Mustafa, türk lehçeleri arasındaki benzer kelimelerin eş değerlik durumu hakkında bir çalışma yapmıştır [1]. Çalışmasında türk lehçelerinin dini, coğrafi, kültürel değişimlerinin kelimeler üzerindeki değişimlerini incelemiştir. Bire bir veya bire çok eş anlamlısı bulunan kelimelerin köklerine inerek tam benzerlik veya kabul edilebilir benzerliklerini göstermiştir. Hasan, yaptığı çalışmada parkinson hastalarının ses öznitelikleri üzerinde çeşitli makine öğrenmesi yöntemleri uygulamıştır [2]. PDC veri seti üzerinde TBA ve DDA boyut indirgeme yöntemi uygulayarak doğruluk değerlerini karşılaştırmıştır. Ahmet rulman arızalarının tespit edilebilmesi için arızalı ve sağlam rulman seslerini toplamış ve fourier dönüşümü kullanarak özniteliklerini çıkarmıştır [3]. Ses sinyallerini on farklı sınıflandırma yöntemi ile gerçekleşmiş ve doğruluk değerlerini karşılaştırmıştır.

III. SES DOSYALARININ TOPLANMASI

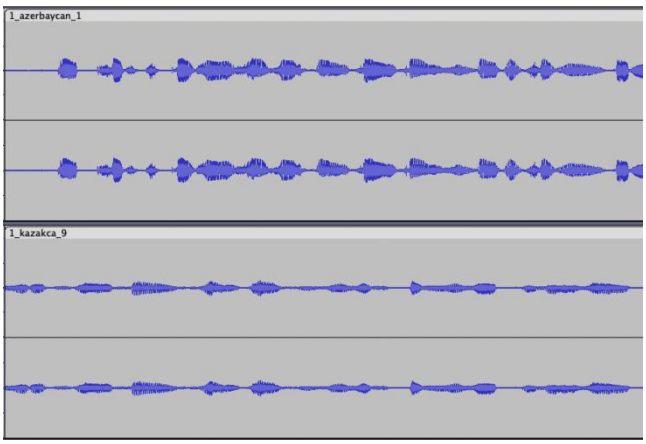
Ses dosyalarının elde edilmesi için öncelikle arama kriterleri belirlenmiştir. Arama kriterleri için sınıflandırma işlemi yapılacak lehçelerin kendi dillerinde 'haber' içeriği taşıyan videolar aranmıştır. Ses ve video için sonsuz kaynağa sahip olan youtube aracı kullanılmıştır. Youtube üzerinde Azerice, Kazakca, Kırgızca, Özbekce, Tatarca, Türkmence, Uygurca dillerinde haber videoları araştırılmış ve spiker sunumu bulunan 20-30 dakikalık video linkleri toplanmıştır. Toplanan linklerin örnek görüntüsü şekil 1'de verilmiştir. Link bağlantıları kaynakların altında url olarak verilmiştir. Ses dosyaları toplama aşamasında sınıflara eşit ses dosyası düşmesi için ses süreleri dikkate alınmıştır. Ses dosyalarının toplanması için toplanan linkler bir dosya ile python dilinde yazılmış olan scripte girdi olarak verilmiştir. Hazırlanan script aracılığı ile linkler "wav" formatına dönüştürülerek yerel dosya hiyerarşisine kayıt edilmiştir. Ham verilerin kayıt aşamasında her bir sınıfa ait dosyalar oluşturulmuştur.

Link	**Lehce**
https://youtu.be/SNxlwSeu2Ng	Azerice
https://youtu.be/I5ijLszpAGY	Tatarca
https://youtu.be/spB2wkwISXE	Turkmence
https://youtu.be/Sfn9CnpvYcs	Uygurca
https://youtu.be/-3A28X5CKw0	Kazakca
https://youtu.be/0yJsBMgV1hM	Kırgızca
https://youtu.be/7q1qGWCW7sk	Özbekce

Şekil 1. Link ve yöreye ait linklerin örnek gösterimi

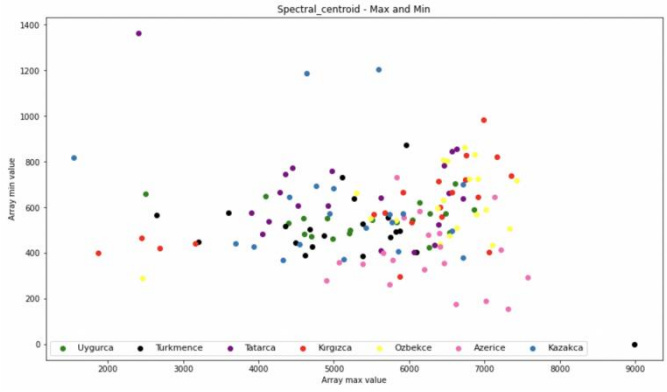
IV. SES DOSYALARININ ANALIZ EDILMESI VE TEMIZLENMESI

Toplanan ses dosyalarından her bir lehçeye ait 10 saniyelik 20'şer adet veri alınmıştır. Alınan ses kesitleri Audacity yazılımı aracılığı ile incelenmiştir. Her bir ses dosyası aynı proje içerisinde açılarak spektrogramları incelenmiştir. Veriler üzerinde bilgi sahibi olduktan sonra seslerin bölümlendirilmesi adımına geçilmiştir. Spektrogram analizinin ekran kesiti şekil 2' de verilmiştir.

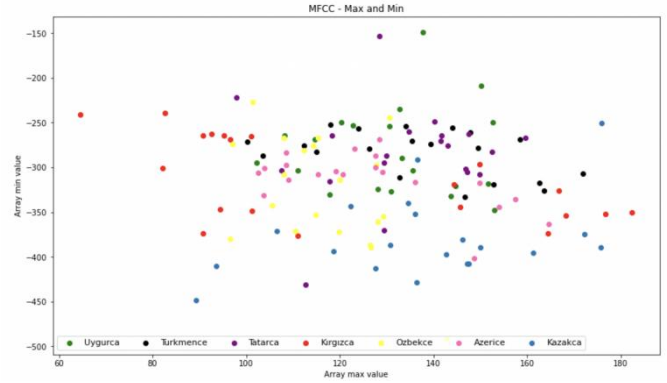


Şekil 2. Audacity yazılımı üzerinde spektrogram analizi

Örnek olarak alınan ses dosyalarını MFCC ve Spectral centroid niteliklerini çıkararak grafik ortalama değerlerini grafik üzerinde gösterdim. Ses dosyaları arasındaki korelasyonu gözlemleyerek veri seti hakkında daha detaylı bilgi edinmiş oldum. Bu çalışma sonucunda grafiğe dökülmüş görseller şekil 3 ve 4 de gösterilmiştir.



şekil 3. Spectral centroid ortalama değerleri üzerinde korelasyon



şekil 4. MFCC ortalama değerleri üzerinde korelasyon

V. SES DOSYALARıNıN BÖLüMLENMESI

Etiketlenmiş ve indirilmiş olan ham ses verileri üzerinde bölümlendirme yapmak gereklidir. Bu işlem için python dilinde script yazılmış ve librosa kütüphanesi kullanılarak 4 saniyelik parçalara ayrılmıştır. Parçalara ayrılan ses dosyaları ve etiketleri csv dosyasına kayıt edilmiştir. Kayıt edilirken kullanılan isim ve sınıf özellikleri korunmuştur. Sınıflara ait dosyalar işaretlenmiş ve tek bir klasör altında toplanmıştır. Kayıt edilirken kullanılan isimlendirme ve dosya yolu şekil 5'de tabloda gösterilmiştir.

file_name	class_name
4_azerbaycan_12.wav	Azerice
2_turkmence_42.wav	Turkmence
1_kirgizca_3.wav	Kırgızca
3_ozbekce_44.wav	Özbekce

Şekil 5. Bölümlenmiş ses dosya örnekleri

Bütün parçalanan ses dosyaları yeniden isimlendirilerek bir klasör altında toplanmıştır. Herbir lehçe için toplanan ses dosyaları sayısı şekil 6'de verilmiştir.

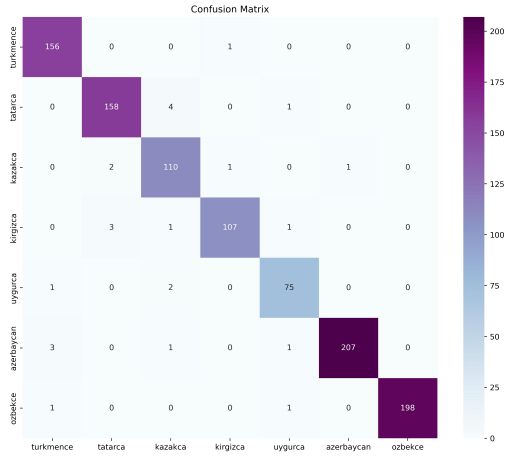
Lehce	Veri adedi	Ses uzunluğu (saniye)
Azerice	369	4
Kazakca	685	4
Kırgızca	524	4
Özbekce	372	4
Tatarca	610	4
Uygurca	1074	4
Türkmence	998	4
Toplam	4632	4

Şekil 6. Ses dosyaları sayısı

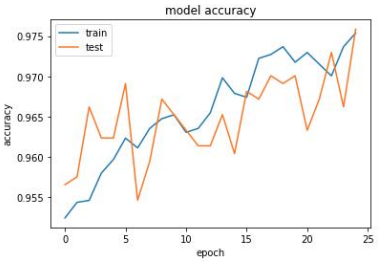
VI. Makine öğrenmesi yöntemleri

1. Evrişimsel sinir ağı (convolutional neural network)

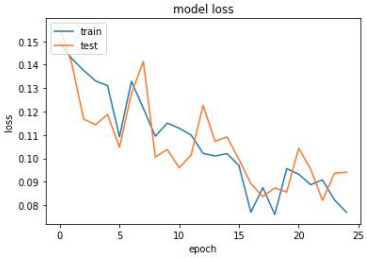
Evrişimsel sinir ağı derin öğrenmenin alt dalıdır. Sınıflandırma ve görüntü işlemlerinde sıklıkla kullanılır. Ses dosyalarını evrişim işlemine sokmadan önce ses dosyalarından nitelik çıkarma işlemleri yapıldı. Bu işlemden sonra her ses dosyası için matris elde edildi. Model için giriş değerleri öznitelik sayısı olarak belirlendi. Model üzerinde katmanlar oluşturuldu ve ezberlemenin önüne geçebilmek için dropout katmanı eklendi. Veri seti %80 eğitim %20 test olarak ayrıldı. Model eğitiminden sonra %97.5 doğruluk, %9.4 kayıp sonuçları elde edildi. Kayıp ve eğitim grafikleri şekil 8-9'da gösterilmiştir. Hata matrisi çıkartıldı. Elde edilen hata matrisi şekil 7'de gösterilmiştir.



Şekil 7. CNN hata matrisi



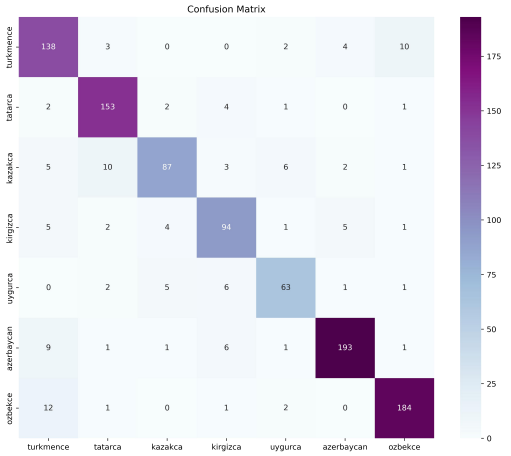
Şekil 8. CNN model doğruluk grafiği



Şekil 9. CNN model kayıp grafiği

2. Destek Vektör Makinesi (Support vector machine)

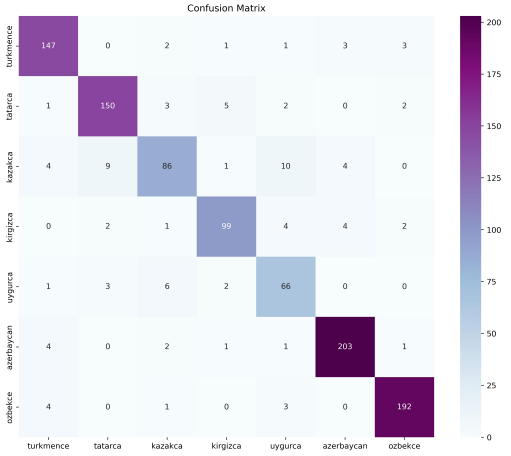
Makine öğrenmesinde , destek vektör makineleri (SVM'ler vektörel ağları destekler), sınıflandırma ve regresyon analizi için kullanılan veriyi analiz eden ilişkili öğrenme algoritmalarıyla denetimli öğrenme modelleridir. Her biri, her iki kategoriden birine ya da diğerine ait olarak işaretlenmiş bir dizi eğitim örneği verildiğinde, bir SVM eğitim algoritması, bir olasılık dışı ikili doğrusal sınıflandırıcı haline getirerek bir kategoriye ya da diğerine yeni örnekler atayan bir model oluşturur. Bu model eğitimi sonucunda %88 doğruluk sonucu elde edildi. Bu işlemden sonra test veri kümesi üzerinde hata matrisi gösterildi. Hata matrisi şekil 10'de gösterilmiştir.



Şekil 10. SVM confusion matris

3. Yapay Sinir Ağı YSA

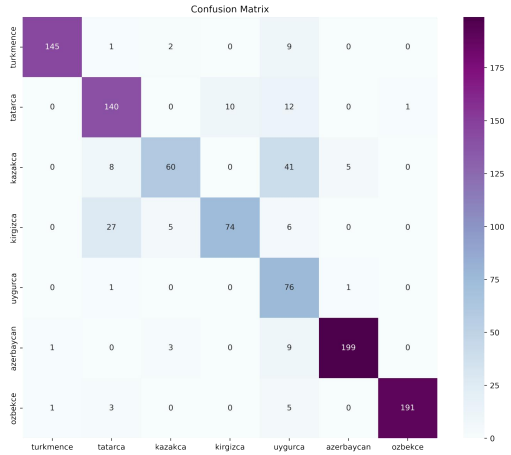
Yapay sinir ağları (YSA), insan beyninin özelliklerinden olan öğrenme yolu ile yeni bilgiler türetebilme, yeni bilgiler oluşturabilme ve keşfedebilme gibi yetenekleri, herhangi bir yardım almadan otomatik olarak gerçekleştirebilmek amacı ile geliştirilen bilgisayar sistemleridir[4]. Yapay sinir ağları insan beyni örnek alınarak, öğrenme sürecinin matematiksel olarak modellenmesi sonucu ortaya çıkmıştır. Beyindeki biyolojik sinir ağlarının yapısını, öğrenme, hatırlama ve genelleme kabiliyetlerini taklit eder[5]. Yapay sinir ağlarında öğrenme işlemi örnekler kullanılarak gerçekleştirilir. Öğrenme esnasında giriş çıkış bilgileri verilerek, kurallar koyulur. Bu model eğitimi sonucunda %95.4 doğruluk sonucu elde edildi. Bu işlemden sonra test veri kümesi üzerinde hata matrisi gösterildi. Hata matrisi şekil 11'de gösterilmiştir.



Şekil 11. YSA Confusion matris

4. Transfer Learning

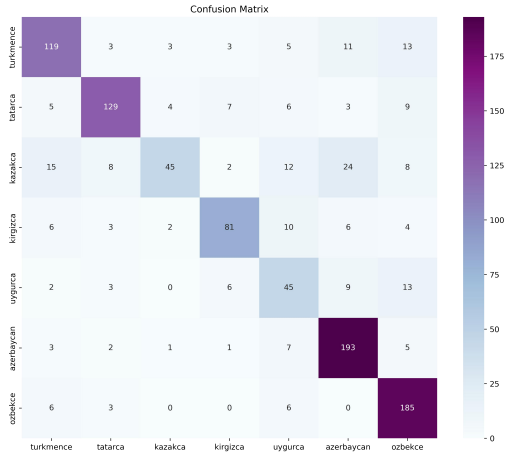
Transfer öğrenimi, yeni bir görevde bir model için başlangıç noktası olarak önceden eğitilmiş bir modeli yeniden kullandığımız bir makine öğrenimi yöntemidir. Basitçe söylemek gerekirse, bir görev üzerinde eğitilmiş bir model, ikinci görevi modellerken hızlı ilerlemeye izin veren bir optimizasyon olarak ikinci bir ilgili görevde yeniden kullanılır. Bu model eğitimi sonucunda %85.4 doğruluk, %43.2 kayıp sonuçları elde edildi. Bu işlemden sonra test veri kümesi üzerinde hata matrisi gösterildi. Hata matrisi şekil 12'de gösterilmiştir.



Şekil 12. Transfer learning Confusion matris

5. Naive Bayes

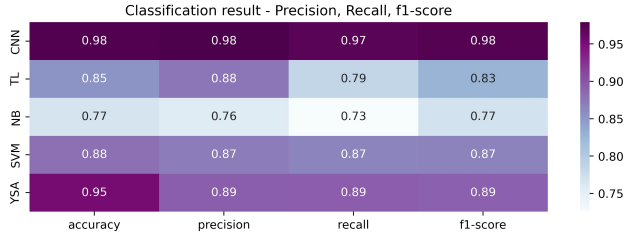
Naïve Bayes sınıflandırıcı, örüntü tanıma problemine ilk bakışta oldukça kısıtlayıcı görülen bir önerme ile kullanılabilen olasılıksal bir yaklaşımdır. Bu önerme, örüntü tanımada kullanılacak her bir tanımlayıcı öznitelik ya da parametrenin istatistik açıdan bağımsız olması gerekliliğidir. Bu model eğitimi sonucunda %76.7 doğruluk sonucu elde edildi. Bu işlemden sonra test veri kümesi üzerinde hata matrisi gösterildi. Hata matrisi şekil 13'de gösterilmiştir.



Şekil 13. Naive Bayes Confusion matris

6. Karşılaştırma

Elde edilen sonuçlar bir matris üzerinde toplandı. Bu matris de modellere ait doğruluk, kesinlik, duyarlılık, f1-skor değerleri gösterilmiştir. Karşılaştırma matrisi şekil 14'de gösterilmiştir.



Şekil 14. Karşılaştırma matrisi

KAYNAKLAR

[1] Uğurlu, Mustafa. "TÜRK LEHÇELERİ ARASINDA BENZER KELİMELERİN EŞ DEĞERLİK DURUMU." *Electronic Turkish Studies* 7.4 (2012).

[2] BADEM, Hasan. "Parkinson Hastaliğinin Ses Sinyalleri Üzerinden Makine Öğrenmesi Teknikleri ile Tanimlanmasi." *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi* 8.2 (2019): 630-637.

[3] TEKTAŞ, Ahmet Burak, and İsmail KIRBAŞ. "RULMAN ARIZALARININ MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE SES ANALİZİ YAPILARAK SINIFLANDIRILMASI."

[4] http://www.ibrahimcayiroglu.com/Dokumanlar/IleriAlgoritmaAnalizi/IleriAlgoritmaAnalizi-5.Hafta-YapaySinirAglari.pdf

[5] https://www.linkedin.com/pulse/yapay-sinir-a%C4%9Flar%C4%B1-ve-tek-katmanl%C4%B1-a%C4%9Flarda-%C3%B6%C4%9Frenme-tanju-do%C4%9Fan/

LINKLER

[LINK]https://www.youtube.com/playlist?list=PLyY6UmEUDM_lwIt1uC0txRo5QHH-Qqqa