**IMDB Üzerinden Duygu Analizi Belirlenmesi**

İsa Kulaksız

**1** Atatürk Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Erzurum, Türkiye (ORCID: 0000-0000-0000-0000)

**Öz**

Son yıllarda dijital ortamlarda artan içerikler (film, dizi, müzik) üzerinde, son tüketicinin film hakkındaki düşüncelerinin olumlu veya olumsuz olma durumunun önemi gitttikçe artmaktadır. 21.yy. da verinin önemi ve özellikle veriden anlamlı çıkarım yapan şirketlerin kar maksimizasyonundaki artıştan dolayı şirketler veri bilimi üzerine ciddi yatırımlar yapmaktadır. Yapılan tüm bu araştırma ve iyileştirmeler şirketlere yaptıkları içeriklerin müşteriler tarafından benimsenip benimsenmediği konusunda fikir vermektedir. Şirketlerde elde ettikleri tüm bu çıktılar üzerinde yapacakları bir sonraki içeriğin akıbeti hakkında bilgi edinmiş olur.

Yazılım dünyasında var olan bu problemin çözümünü, operatör bağımlılığını ortadan kaldırıp teşhis doğruluğunu artırmak için bir karar destek sistemi (DSS) tasarlandı.

1. Giriş

Üzerinde işlem yapılacak olan veri seti 2 adet sütundan oluşmaktadır. Bunlar; kullanıcı yorumu(text), yapılan yorumun pozitif veya negatif durumudur(label)

Veri seti içerisinde yer alan label sınıfı 2 adet parametreye sahiptir. Bunlar; 0 yorumun negatif, 1 pozitif olma durumudur.

A picture containing bubble chart

Description automatically generated

Resim 1.1 IMDB

Kullanılacak öğrenme tipi (Makine öğrenimi) verinin sisteme nasıl sunulduğuna dayanır. Makine öğrenimi algoritması, çoğunlukla her zaman yapılandırılmış veriye ihtiyaç duyarken, Derin öğrenme ağları yapay sinir ağlarının katmanlarına dayanır.[1]

Makine öğrenimi, türettikleri verilere göre öğrenen ya da performansı iyileştiren sistemler oluşturmaya odaklanan bir AI sistemidir. Bu yapılan araştırmada 2 farklı Makine öğrenimi algoritması kullanılmıştır. Bunlar; Lojistik Regresyon ve Karar Ağaçları dır.

Lojistik Regresyon, bir veri setinin önceki gözlemlerine dayalı olarak evet veya hayır gibi gibi ikili bir sonucu tahmin etmeye yönelik istatistiksel bir analiz yöntemidir. Buna ek olarak bir lojistik regresyon modeli, bir veya daha fazla mevcut bağımsız değişken arasındaki ilişkiyi analiz ederek bir bağımlı veri değişkenini tahmin etmeye çalışır.

Karar Ağaçları ise en popüler sınıflandırma algoritmalarından biridir. Bu algoritmada her bir dahili düğümün bir öznitelik üzerinde bir testi ifade ettiği, her dalın testin bir sonucunu temsil ettiği ve her yaprak düğümün bir sınıf etiketi taşıdığı bir ağaç yapısıdır.

Diagram

Description automatically generated

Resim1.2 Örnek bir Karar Ağacı Şeması

Duygu analizi üzerine yapılan çalışmalarıdan Hibrit Özellik Çıkarma Yöntemi (Hybrid Feature Extraction Method) yöntemi özelliklerden özellikleri çıkarmak için kullanılır. HFEM bireysel özellik çıkarma yönteminden çok daha verimli hale geldiği kanıtlanmıştır.[2]

Derin öğrenme üzerinde yapılan çalışmalarda kullanılan CNN tabanlı LSTM modeli IMDB veri setinde duygu polaritesini klasik makine öğrenimi algoritmalarıtla benzer doğruluk tahmin ettiği kanıtlanmıştır.[3]

2. Materyal ve Metot

**2.1. Veri Ön İşleme**

Veri setinde yer alan yorumların olumlu, olumsuz olarak gösterimi Görsel 1’de yer almaktadır. Yapılan yorumların %50,1 olumluyken %49.9’u olumsuz olarak yer almıştır.

*Chart, pie chart

Description automatically generated*

*Görsel 1 Veri seti yapısı*

Sklearn kütüphanesi yardımıyla veri seti test ve eğitim olarak iki sınıfa ayrıldı. Bölme işlemi gerçekleştikten sonra eğitim ve test boyutları Tablo 1’de gösterilmiştir.

|  |  |
| --- | --- |
| X | Boyut (Shape) |
| Original | (39723, 1) (39723,) |
| X\_train | (31778, 119535) (31778,) |
| X\_test | (7945, 119535) (7945,) |

*Tablo 1 X*

***2.1.1. Model’in Hazırlanması***

Diagram

Description automatically generated

Veri setinde tekrar eden en çok olumlu ve olumsuz kelimeler wordcloud üzerinden hesaplanmıştır. İlgili görsel Görsel 2’de yer almaktadır. Burada dikkat edilen kelimeler incelendiğinde “bad, although vb.” olumsuz kelime tekrarının fazla olduğu kanıtlanmıştır.



*Görsel 2 En çok tekrar eden olumsuz kelimeler*

3. Araştırma Sonuçları ve Tartışma

**3.1. Araştırma Sonuçları**

*Tablo 2. Doğruluk (Accuracy) oranları*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ML Algoritma | Accuracy (Doğruluk) | Precision | Recall | F1-Score |
| Logistic Regression | %89,08 | %87,48 | %90,49 | %88,96 |
| Decision Tree | %84,53 | %86,11 | %81,27 | %83,62 |

Tablo 2’de yer alan parametreler kullanılarak model eğitildikten sonra matplotlib kütüphanesi kullanılarak accuracy, precision, recall, f1-score değeri hesaplanmaktadır.

Diagram, table

Description automatically generatedA picture containing text

Description automatically generated

*Confusion Matrix*

Doğruluk (accuracy) bir modelin başarısını ölçmek için kullanılan önemli bir metriktir.[4]

Elde edilen sonuçlarda dikkat edilmesi gereken diğer bir önemli noktaysa ROC eğrisi ve altında kalan kısımdır. ROC (Receiver Operating Characteristic) = TPR / FPR temelde sınıflandırma problemlerini çözmek, modellerin iyi çalışıp çalışmadığını kontrol eden bir metriktir. Görsel 3 ve 4’te yer alan görseller üzerinde temsil edilmektedir. Her iki görselde de ROC eğrilerinin altında kalan alan 1’e yaklaştıkça sınıflandırma için kurulan modelin başarımı arttırdığı kanıtlanmıştır.

Bu alanda yapılan çalışmalarda farklı pek çok yöntem izlenmiştir. Derin öğrenme CNN LSTM tabanlı yöntemden elde edilen doğruluk %89’dur.

Makine öğrenimi ile yapılan araştırmaların Derin Öğrenme CNN tabanlı LSTM yönteminden ciddi bir artış yapmadığı kanıtlanmıştır.

Chart, line chart

Description automatically generated*Chart, line chart

Description automatically generated*

*Görsel 3 Logistic Regression ROC Eğrisi. Görsel 4 Decision Tree ROC Eğrisi*

4. Tartışma

Proje üzerinde en yüksek sonuç veren Makine Öğrenimi algoritması Lojistik Regresyondur. Bu algoritmada parametre optimizasyonu çok düşükte olsa doğruluk değerinin artışı gözlenebilir.

Elde edilen sonuçlar içerik geliştiricilerle paylaşılmalı ve sonuçlara göre yeni bir içeriğin nasıl olması gerektiğine dair önemli bir fikir verebilir.

Kaynakça

[1] B. I. Academy, “DEEP LEARNING & MACHINE LEARNING,” *Bilginç IT Academy*. https://bilginc.com/tr/blog/5427/deep-learning-and-machine-learning (accessed May 28, 2022).

[2] H. M. Kumar, B. S. Harish, and H. K. Darshan, “Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method.,” *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 5, 2019.

[3] A. Yenter and A. Verma, “Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis,” in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, 2017, pp. 540–546.

[4] G. Öğündür, “Doğruluk (Accuracy) , Kesinlik(Precision) , Duyarlılık(Recall) ya da F1 Score ?,” *Medium*, Apr. 22, 2020. https://medium.com/@gulcanogundur/do%C4%9Fruluk-accuracy-kesinlik-precision-duyarl%C4%B1l%C4%B1k-recall-ya-da-f1-score-300c925feb38 (accessed May 28, 2022).