IMDb Film İnceleme Duygu Analizi

Muhammed Çağrı Öz-Eray Naldöken
Elektronik Elektronik Mühendisliği
TOBB Ekonomi ve Teknoloji Üniversitesi
muhammedcagrioz@etu.edu.tr-enaldoken@etu.edu.tr

Özet- Bu proje, IMDB veri setindeki film yorumlarını olumlu ya da olumsuz duygular olarak sınıflandırmayı amaçlamaktadır. Veri seti 50.000 yorumu içermektedir ve görev, veri temizleme, keşifsel veri analizi (EDA) ve çeşitli tahmin modellerinin uygulanmasını içeren kapsamlı bir yaklaşım gerektirmektedir. Projede, Lineer Regresyon, Random Forest ve Lojistik Regresyon makine öğrenimi algoritmaları kullanılmıştır. Lojistik regresyon ile yaklaşık %89 maksimum doğruluk oranına ulaşmıştır. Bu çok yönlü yaklaşım, duygu sınıflandırması için farklı yöntemlerin analiz edilmesini ve karşılaştırılmasını sağlamaktadır.

Anahtar Kelimeler – Regresyon, doğruluk, makine öğrenme.

I. GİRİŞ

Duygu analizi, pazar analizi, müşteri davranışı, ürün incelemeleri ve sosyal medya takibi gibi geniş bir uygulama yelpazesi nedeniyle doğal dil işleme (NLP) alanında önemli bir yer kazanmıştır. Metin tabanlı verilerde temsil edilen duyguları doğru bir şekilde sınıflandırmak, işletmelerin, şirketlerin ve araştırmacıların müşteri davranışları ve kamuoyu hakkında önemli bilgiler edinmelerini sağlar. 50.000 yorumu içeren IMDb veri seti, duygu analizi algoritmalarını değerlendirmek için bir standart haline gelmiştir ve pozitif ya da negatif olarak sınıflandırılması gereken zorlayıcı ve çeşitli bir inceleme koleksiyonu sunmaktadır.

Bu çalışma, IMDb veri seti kullanılarak yapılan bir duygu sınıflandırması analizini sunmaktadır. Proje, film yorumlarını sınıflandırmada çeşitli makine öğrenimi metodlarını kullanıp bunları kıyaslamaktadır. Makine öğrenimi yöntemleri, duygu analizi görevlerinde basit ama etkili yaklaşımlar sundukları için yaygın olarak kullanılan Lineer Regresyon, Random Forest ve Lojistik Regresyon gibi teknikleri içermektedir.

Bu çalışmanın birincil amacı, duygu sınıflandırmasında doğruluk ve sağlamlık açısından farklı modellerin etkilerini ve performanslarını karşılaştırmaktır. Bu karşılaştırma aracılığıyla çalışmayı anlamlandırmak, her yaklaşımın güçlü ve zayıf yönleri hakkında içgörüler sağlamak ve her bir modelin hangi koşullarda başarılı olduğunu vurgulamayı hedeflemektedir.

.

II. LİTERATÜR TARAMASI

Özellikle lineer regresyon, random forest ve lojistik regresyon gibi yöntemler, duygu analizinde daha iyi performans sergileyen teknikler olarak öne çıkmaktadır. Reddy, R., & Kumar [2], lojistik regresyonun film incelemeleri üzerindeki duygu analizinde yüksek doğruluk oranları sağladığını belirtmişlerdir. Ayrıca, Shah, K., Patel, H., Sanghvi, D., & Shah, M.,[3] lojistik regresyon ve random forest algoritmalarının duygu sınıflandırmasında önemli başarılara ulaştığını, ancak lojistik regresyonun en yüksek doğruluk oranını elde ettiğini vurgulamaktadır. Lojistik regresyon, metin verisi üzerindeki ikili sınıflandırma problemlerinde özellikle etkin sonuçlar vermektedir. **G**eorge B. Aliman[4], lojistik regresyon modelinin duygu analizi için diğer modellerle karşılaştırıldığında daha hızlı ve etkili sonuçlar sunduğunu bulmuşlardır. Muralidhar, A., & Lakkanna, Y. [5] ise, duygu analizinde kullanılan farklı makine öğrenimi yöntemlerini karşılaştırarak, modeller arasında lojistik regresyonun en iyi doğruluğu sağladığını ortaya koymuşlardır. Random Forest ve lineer regresyon gibi diğer modeller de başarılı sonuçlar elde etmiş, ancak lojistik regresyon, film incelemelerinin duygu analizi gibi doğal dil işleme (NLP) uygulamalarında daha yüksek doğruluk oranları ile öne çıkmıştır. Couronn&eacute, R., Probst, P., & Boulesteix [6], random forest ve lojistik regresyon algoritmalarının ikili sınıflandırma problemlerindeki performanslarını karşılaştırmış ve lojistik regresyonun en iyi performansı gösterdiğini tespit etmişlerdir.

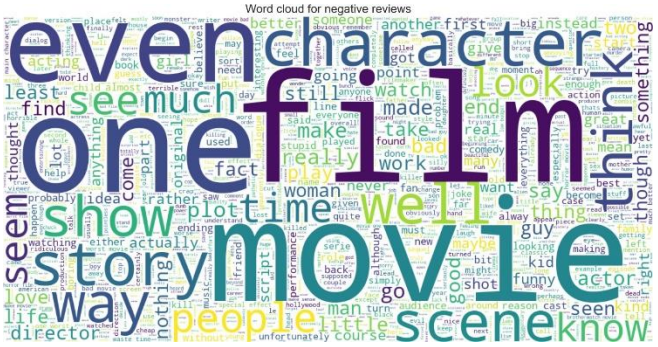
Sonuç olarak, literatür, duygu analizinde doğrusal ve ağaç tabanlı modellerin etkinliğini vurgulamaktadır. Özellikle lojistik regresyon, film incelemeleri ve benzeri metin verisi üzerinde üstün doğruluk ve performans elde etme potansiyeli ile tercih edilen model haline gelmiştir. Kamath, C. N., Bukhari, S. S., Dengel, A.,**[7]**, ileri sinir ağları ve derin öğrenme tekniklerinin de giderek daha fazla tercih edilse de, geleneksel makine öğrenimi modellerinin özellikle de ham verilerle hala etkili olduğunu ve genellikle daha hızlı sonuçlar sunduğunu belirtmektedir.

III. SEÇİLEN VERİ SETİ VE ÖN İŞLEME

A. Veri Seti

Bu çalışmada, ikili duygu sınıflandırması için etiketlenmiş 50.000 film incelemesinden oluşan IMDb film incelemeleri veri seti kullanılmıştır. Veri seti, iki sütundan oluşmaktadır: "review" sütunu, film incelemesinin metnini içerirken, "sentiment" sütunu incelemede ifade edilen olumlu veya olumsuz duygu durumunu belirtir. Bu veri seti, doğal dil işleme (NLP) görevlerinde yaygın olarak kullanılmaktadır.

Şekil 1. Olumlu Yorumlar İçin Kelime Bulutu.



Şekil 2. Olumsuz Yorumlar İçin Kelime Bulutu.

B. Keşifsel Veri Analizi (EDA)

EDA kapsamında veri setine dair içgörüler elde etmek için çeşitli analizler gerçekleştirilmiştir. İlk olarak, olumlu ve olumsuz incelemeler arasındaki dengeyi sağlamak için duygu etiketlerinin dağılımı detaylı incelenmiştir. Ayrıca, inceleme uzunluklarının dağılımı incelenerek herhangi bir önemli varyasyon olup olmadığı tespit edilmiştir. Her bir duygu sınıfıyla ilişkili en sık kullanılan terimleri görselleştirmek amacıyla, olumlu ve olumsuz incelemelerde yer alan en sık geçen kelimeleri gösteren kelime bulutları oluşturulmuştur.

C. Ön İşleme

Veri setini modellerin girişi olarak kullanılabilir hale getirmek için kapsamlı bir ön işleme aşaması uygulanmıştır. Bu ön işleme adımları şunları içermektedir:

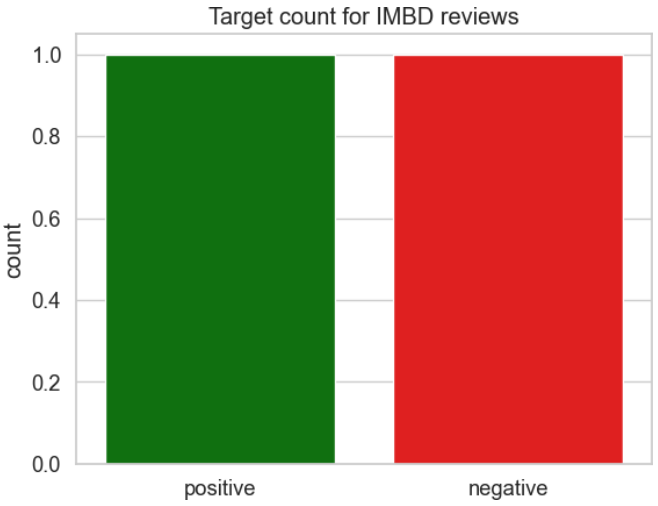
• Metin Temizleme: BeautifulSoup kütüphanesi kullanılarak HTML etiketleri kaldırılmıştır. Böylece metin verisi, model performansını etkileyebilecek işaretlemelerden arındırılmıştır. Ayrıca, tüm metinler küçük harfe dönüştürülmüştür.

• Tokenizasyon: NLTK kütüphanesindeki word_tokenize fonksiyonu kullanılarak her bir inceleme, ayrı kelimelere bölünmüştür. Bu adım, metni daha küçük ve işlenebilir parçalara ayırmak için önemlidir.

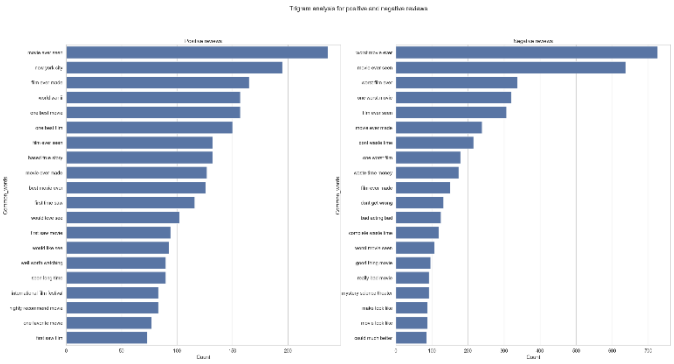
• Stopwords (Yardımcı Kelimeler) Kaldırma: İngilizce'de sıkça kullanılan ve duyguya anlamlı katkı sağlamayan (örneğin, "a", "an", "the" gibi) yaygın kelimeler, NLTK'nin stopwords kütüphanesi kullanılarak kaldırılmıştır.

• Lemmatizasyon: Kelimeler, WordNetLemmatizer fonksiyonu kullanılarak kök ya da temel formlarına dönüştürülmüştür. Bu adım, aynı kelimenin farklı biçimlerini aynı şekilde ele alarak metni daha tutarlı hale getirmiştir.

• Vektörleştirme: Metin verisini sayısal formata dönüştürmek için Count Vectorization (metni, kelime sayılarına dayalı bir matris haline dönüştürür) ve TF-IDF (Terim Frekansı-Ters Belge Frekansı; bir kelimenin bir belge içerisindeki önemini, tüm belgeler arasındaki sıklığına göre ağırlıklandırır) yöntemleri kullanılmıştır.



Şekil 3. IMDb İncelemeleri için Hedef Sayısı



Şekil 4. Triagram Analizi.

Bu ön işleme adımları, IMDb veri setini makine öğrenimi ve derin öğrenme modellerine uygun hale getirmek ve metin verisini doğru duygu sınıflandırması için uygun bir formata dönüştürmek için gerekliydi.

.

IV. SEÇİLEN MODELLER

A. Kullanılan Modellerin Genel Bakışı

Bu çalışmada, IMDb veri setinde film incelemelerinin duygu sınıflandırılması için çeşitli makine öğrenimi modelleri kullanılmıştır. Seçilen modeller şunlardır:

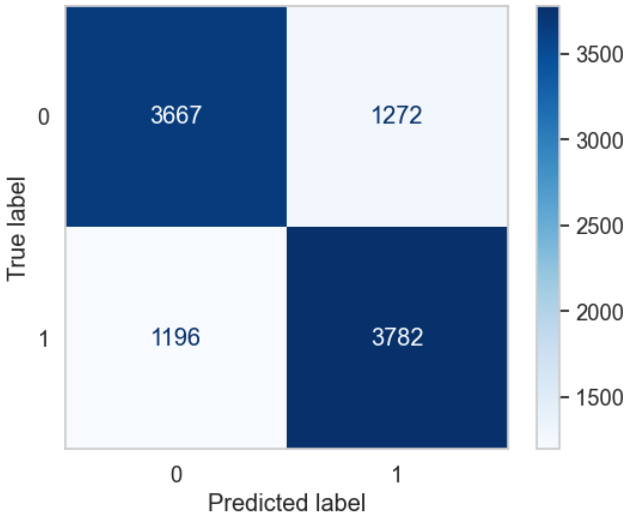
- **Lineer Regresyon:** Basit fakat etkili bir doğrusal model olup, sürekli değişkenler arasında ilişkiyi modellemek için kullanılır. İkili sınıflandırma görevlerinde sınıf olasılıklarını tahmin etmek için de uygulanabilir.
- **Random Forest:** Birden fazla karar ağacından oluşan topluluk tabanlı bir modeldir. Karar ağaçlarının bir araya gelmesiyle daha sağlam tahminler elde edilmesini sağlar ve genellikle overfitting (aşırı öğrenme) sorununu minimize eder.
- **Lojistik Regresyon:** Basit fakat etkili bir doğrusal model olup, ikili sınıflandırma için kullanılır. Bir girdinin belirli bir sınıfa ait olma olasılığını tahmin eder. Genellikle doğru sınıflandırmalar için verimli ve hızlı sonuçlar verir.

B. Model Performansları

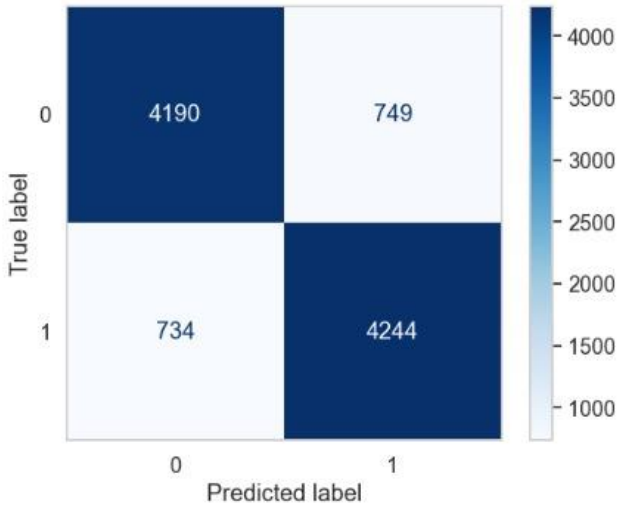
Veri seti büyük olmasına rağmen, Lineer Regresyon, Lojistik Regresyon (LR) ve Random Forest (RF) gibi geleneksel makine öğrenimi yöntemlerinin, özellikle karmaşık ve büyük veri setlerinde zayıf sonuçlar vermesi beklenirken, sonuçlar şaşırtıcı derecede iyi olmuştur. Bunun nedeni, veri setinin belirli bir düzeyde doğrusal ilişkilere sahip olması ve modelin başarılı bir şekilde parametrelerinin optimize edilmiş olması olabilir. Ayrıca, Random Forest modelinin yüksek doğruluk ve sağlamlık göstermesi, karar ağaçlarının çeşitliliği ve topluluk yapısının gücünden kaynaklanmaktadır.

V. SONUÇLAR

Bu çalışmada, IMDb film inceleme veri seti üzerinde duygu analizi için klasik makine öğrenimi yöntemlerini değerlendirdik ve elde edilen sonuçları tartıştık. Üzerinde çalıştığımız makalede Yapay Sinir Ağları (Artificial Neural Networks- ANN) yöntemi kullanılırken, bu çalışmada kullanılan modeller sırasıyla Lineer Regresyon (LR), Random Forest (RF) ve Lojistik Regresyon (LR) olmuştur.



Şekil 5. Lineer Regresyon Sınıflandırıcı Karışıklık Matrisi.



Şekil 6. Random Forest Sınıflandırıcı Karışıklık Matrisi.

Çalışmamızın sonuçlarına göre, Lojistik Regresyon (LR) modeli en yüksek doğruluk oranını (%89) elde ederek diğer yöntemlerin önüne geçmiştir. Random Forest, veri setinin özelliklerini etkili bir şekilde değerlendirmesiyle iyi bir performans sergilerken, Lineer Regresyon'un doğruluk oranı beklenenden düşük kalmıştır. Bu sonuçlar, Lojistik Regresyon'un ikili sınıflandırma görevleri için sağlam bir seçenek olduğunu bir kez daha doğrularken, veri setinin yapısına ve ön işleme adımlarının etkinliğine de dikkat çekmektedir.

```
Lojistik Regresyon Doğruluğu: 89.11%
Lojistik Regresyon Metrikleri:

              precision    recall  f1-score      support
0              0.903408  0.874873  0.888912  4939.000000
1              0.879626  0.907192  0.893196  4978.000000
accuracy       0.891096  0.891096  0.891096     0.891096
macro avg      0.891517  0.891033  0.891054  9917.000000
weighted avg   0.891470  0.891096  0.891062  9917.000000
```

Şekil 7. Lojistik Regresyon Metrikleri

Yapay Sinir Ağları (ANN) gibi daha karmaşık ve hesaplama maliyeti yüksek yöntemlerle karşılaştırıldığında, basit modellerin daha az kaynak tüketirken rekabetçi bir doğruluk seviyesi sunabileceği gözlemlenmiştir. Bu durum, veri setinin iyi yapılandırılmış ve belirli bir ölçüde doğrusal ilişkilere sahip olmasıyla açıklanabilir. Bu çalışma, veri analizi ve model seçimi sürecinde daha basit yöntemlerin, doğru senaryolarda güçlü bir alternatif olabileceğini göstermiştir. Özellikle Lojistik Regresyon gibi modellerin, uygun ön işleme ile yüksek doğruluk oranlarına ulaşabilmesi, makine öğrenimi modelleri arasında seçim yaparken dikkat edilmesi gereken önemli bir faktördür.

KAYNAKLAR

[1] M. King, B. Zhu, and S. Tang, “Optimal path planning,” *Mobile Robots*, vol. 8, no. 2, pp. 520-531, March 2001.

[2] Reddy, R., & Kumar, U. M. A. (2022, July 6). Classification of user’s review using modified logistic regression technique- international journal of system assurance engineering and management.

[3] Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020, March 5). A comparative analysis of logistic regression, Random Forest and KNN models for the text classification - augmented human research.

[4] Sentiment analysis using logistic regression. (n.d.).

[5] Muralidhar, A., & Lakkanna, Y. (2024). *Machine Learning Models for Sentiment Analysis*. ieeexplore. https://ieeexplore.ieee.org/abstract/document/10717071/authors#authors

[6] Couronné, R., Probst, P., & Boulesteix, A.-L. (2018, July 17). *Random Forest versus logistic regression: A large-scale benchmark experiment - BMC Bioinformatics*.

[7] Kamath, C. N., Bukhari, S. S., Dengel, A., Cannannore Nidhi KamathGerman Research Center for Artificial Intelligence (DFKI), U. of K. P., Syed Saqib BukhariGerman Research Center for Artificial Intelligence (DFKI), U. of K. P., & Andreas DengelGerman Research Center for Artificial Intelligence (DFKI), U. of K. P. (2018, August 28). *Comparative study between traditional machine learning and deep learning approaches for text classification: Proceedings of the ACM symposium on document engineering 2018*. ACM Conferences. https://dl.acm.org/doi/abs/10.1145/3209280.3209526