LSTM TABANLI DUYGU ANALİZİ

Ali Bora VARİNLİ
22502405022@subu.edu.tr

Özet

Bu çalışmada, IMDb (Internet Movie Database) veri kümesi kullanılarak bir duygu analizi modeli geliştirildi. IMDb, sinema endüstrisine ait geniş bir veritabanı sunan ve dünya çapında popüler olan bir çevrimiçi film ve dizi inceleme platformudur. IMDb veri kümesi, kullanıcıların filmler ve diziler hakkındaki incelemelerini içeren büyük bir metin koleksiyonunu barındırır. Bu çalışmada, LSTM tabanlı bir yapay sinir ağı modeli kullanıldı ve model IMDb veri kümesi üzerinde filmler hakkındaki duygusal tepkileri tahmin etmek için eğitildi. Modelin performansı değerlendirildi ve filmler hakkındaki duygusal tepkileri doğru bir şekilde tahmin edebilme yeteneği ortaya konuldu. Bu çalışma, işletmelerin pazarlama stratejilerini optimize etme, kullanıcı deneyimlerini iyileştirme ve sosyal medya etkileşimlerini yönlendirme gibi alanlarda değerli bir araç olarak kullanılabilecek bir duygu analizi modelinin geliştirilmesini amaçlamaktadır.

Abstract

In this study, a sentiment analysis model was developed using the IMDb (Internet Movie Database) dataset. IMDb is a globally popular online movie and series review platform that offers a large database of the movie industry. The IMDb dataset contains a large collection of texts that contain reviews from users about movies and TV shows. In this study, an LSTM-based neural network model was used and the model was trained to predict emotional responses about movies on the IMDb dataset. The performance of the model was evaluated and its ability to accurately predict emotional reactions about movies was demonstrated. This study aims to develop a sentiment analysis model that can be used as a valuable tool in areas such as optimizing the marketing strategies of businesses, improving user experiences and directing social media interactions.

1. Giriş

Duygu analizi, metin verilerinin incelenmesi ve içerdikleri duygusal anlamların belirlenmesi için kullanılan bir yapay zeka alt alanıdır. Metin tabanlı duygu analizi, insanların sosyal medya paylaşımları, ürün incelemeleri, haberler ve diğer çevrimiçi içerikler aracılığıyla ifade ettikleri duygusal tepkileri anlamak için yaygın olarak kullanılan bir yöntemdir [1]. Bu alandaki çalışmalar, işletmelerin müşteri memnuniyetini değerlendirmek, kamusal tepkileri izlemek ve sosyal medya kampanyalarını optimize etmek gibi birçok uygulama alanında büyük önem taşımaktadır.

Bu çalışmanın amacı, IMDb (Internet Movie Database) veri kümesi kullanılarak bir duygu analizi modeli geliştirmektir. IMDb, sinema endüstrisine ilişkin geniş bir veritabanı sunan ve dünya çapında popüler olan bir çevrimiçi film ve dizi inceleme platformudur [2]. IMDb veri kümesi, kullanıcıların filmler ve diziler hakkındaki incelemelerini içeren büyük bir metin koleksiyonunu barındırmaktadır. Bu incelemeler, kullanıcıların filmler hakkındaki görüşlerini, eleştirilerini ve duygusal tepkilerini yansıtmaktadır [3]. Bu zengin veri kaynağı, duygu analizi modellerinin geliştirilmesi için değerli bir kaynak haline gelmiştir.

Bu çalışmada, IMDb veri kümesi kullanılarak bir duygu analizi modeli oluşturulmuştur. İlk adımda, veri kümesi yüklenmiş ve eğitim ve test veri setleri oluşturulmuştur. Veri kümesi, en sık kullanılan 10,000 kelimeyle sınırlanmış ve her bir inceleme 200 kelimeyle sınırlanmıştır. Ardından, uzun kısa süreli bellek (LSTM) tabanlı bir yapay sinir ağı modeli geliştirilmiştir. LSTM, metin verilerinin zaman bağımlılığını ve bağlamsal ilişkilerini dikkate alabilen güçlü bir modeldir [4]. Model, bir Gömme (Embedding) katmanı, bir LSTM katmanı ve bir yoğunluk (Dense) katmanı içermektedir. Model, filmler hakkındaki duygusal tepkileri tahmin etmek için eğitilmektedir [5].

Bu çalışmanın temel amacı, IMDb veri kümesini kullanarak oluşturulan duygu analizi modelinin performansını değerlendirmek ve filmler hakkındaki duygusal tepkileri doğru bir şekilde tahmin edebilme yeteneğini ortaya koymaktır. Bu model, işletmelerin pazarlama stratejilerini optimize etme, kullanıcıların deneyimlerini iyileştirme ve sosyal medya etkileşimlerini yönlendirme gibi birçok alanda değerli bir araç olarak kullanılabilir.

Duygu analizi alanındaki önceki çalışmalar, genellikle metin verilerini işlemek ve duygusal anlamları çıkarmak için farklı yöntemler ve algoritmalar kullanmaktadır. Ancak, LSTM gibi derin öğrenme modelleri, metinlerdeki bağlamsal ilişkileri daha iyi anlamak ve daha hassas duygu tahminleri yapmak için avantaj sağlamaktadır. Bu çalışma, LSTM tabanlı bir modelin IMDb veri kümesi üzerinde duygu analizi yapabilme yeteneğini göstererek, derin öğrenme yöntemlerinin duygu analizi alanında ne kadar etkili olabileceğini göstermektedir.

Bununla birlikte, duygu analizi modelleriyle ilgili bazı zorluklar da vardır. Metinlerdeki duygusal ifadelerin çok çeşitli olması, yanlış anlamaların ve hatalı tahminlerin ortaya çıkmasına neden olabilir. Ayrıca, duygusal ifadelerin kültürel ve dil bağımlı olması da doğruluk oranını etkileyebilir. Bu nedenle, duygu analizi modellerinin çeşitli veri kaynakları ve dil toplulukları üzerinde test edilmesi ve genelleme yeteneklerinin iyileştirilmesi önemlidir.

2. Materyal ve Metot

2.1. Veri Kümesi

Bu çalışmada, IMDb (Internet Movie Database) veri kümesi kullanılmıştır. IMDb, çevrimiçi bir film ve dizi inceleme platformudur ve kullanıcıların film ve dizi hakkındaki görüşlerini, eleştirilerini ve puanlarını içeren geniş bir veritabanına sahiptir. Veri kümesi, çeşitli filmler hakkındaki incelemelerin metinlerini içermektedir. Bu metinler, kullanıcıların duygusal tepkilerini yansıtmaktadır ve pozitif veya negatif bir duygu ifade edebilir. IMDb veri kümesi, çeşitli film türlerinden ve farklı dil ve kültürlere ait metinleri içeren geniş bir koleksiyona sahiptir.

2.2. Veri Ön İşleme

Veri kümesi, TensorFlow Keras kütüphanesinin sağladığı imdb.load_data() fonksiyonu kullanılarak yüklenmiştir [1]. Bu fonksiyon, veri kümesini eğitim ve test veri setleri olarak ayırır ve en sık kullanılan kelimelerin belirlenmesi için bir kelime indeksi oluşturur. Bu çalışmada, en sık kullanılan 10,000 kelime kullanılmıştır [1].

Veri ön işleme aşamasında, incelemeler belirli bir uzunluğa getirilmiş veya tamamlanmıştır. Her bir inceleme, maksimum 200 kelimeyle sınırlanmıştır. Eğer inceleme 200 kelimeye ulaşmazsa, eksik kalan kısımlar sıfırlarla doldurulmuştur. Veri ön işleme adımları, sequence.pad_sequences() fonksiyonu kullanılarak gerçekleştirilmiştir [1].

2.3. LSTM Modeli

Bu çalışmada, uzun kısa süreli bellek (LSTM) tabanlı bir yapay sinir ağı modeli kullanılmıştır. LSTM, metin verilerindeki zaman bağımlılığını ve bağlamsal ilişkileri dikkate alabilen güçlü bir modeldir [4]. Model, filmler hakkındaki duygusal tepkileri tahmin etmek için eğitilmiştir.
LSTM modeli, TensorFlow Keras kütüphanesinin sunduğu Sequential model kullanılarak oluşturulmuştur [1]. Modelin ilk katmanı, Gömme (Embedding) katmanıdır. Bu katman, kelime indekslerini vektörlerle temsil eder ve kelime dağılımını yakalar. Gömme katmanı, en sık kullanılan 10,000 kelime için 128 boyutlu vektörler oluşturur [1].
LSTM katmanı, Gömme katmanını takip eder. Bu katman, önceki durumlarını hatırlayabilen hücreler kullanarak metinlerdeki bağlamsal ilişkileri düzeyde analiz etme yeteneğine sahiptir. LSTM katmanında 128 hücre kullanılmıştır ve aşırı uyum (overfitting) önlemek için %20 dropout ve tekrarlayan dropout uygulanmıştır [1].
Son katman olarak, yoğunluk (Dense) katmanı kullanılmıştır. Bu katman, LSTM katmanının çıktılarını tek bir çıktı düğümüne indirger ve sigmoid aktivasyon fonksiyonunu kullanarak bir olasılık değeri üretir. Bu olasılık değeri, incelemenin pozitif veya negatif bir duygu ifade etme olasılığını temsil eder [1].

3. Deneysel Çalışma ve Sonuçları

Eğitim aşamasında, IMDb veri kümesinin eğitim veri seti kullanılarak modelin parametreleri optimize edilmiştir. Model, 5 epoch boyunca eğitilmiştir ve her epoch için eğitim veri setindeki kayıp değeri ve doğruluk oranı kaydedilmiştir. Ayrıca, ayrılmış olan test veri seti üzerinde de modelin performansı ölçülmüştür.

Eğitim sürecinde elde edilen sonuçlar şu şekildedir:

Toplam eğitim dizisi sayısı: 25,000
Toplam test dizisi sayısı: 25,000
Eğitim dizilerinin şekli: (25,000, 200)
Test dizilerinin şekli: (25,000, 200)
Eğitim sırasında her bir epoch için kayıp değeri (loss) ve doğruluk oranı (accuracy) aşağıdaki gibidir:

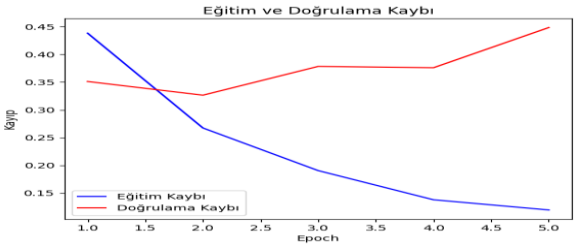
Epoch 1/5:
Eğitim kayıp değeri: 0.4383
Eğitim doğruluk oranı: 0.7943
Doğrulama kayıp değeri: 0.3512
Doğrulama doğruluk oranı: 0.8562

Epoch 2/5:
Eğitim kayıp değeri: 0.2673
Eğitim doğruluk oranı: 0.8936
Doğrulama kayıp değeri: 0.3265
Doğrulama doğruluk oranı: 0.8635

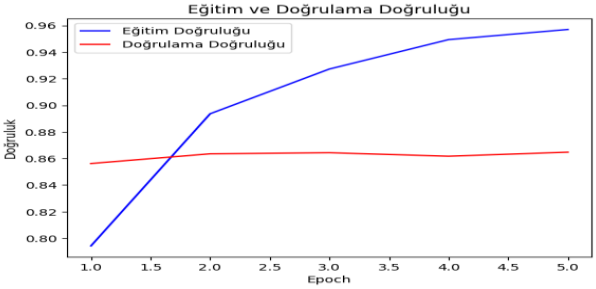
Epoch 3/5:
Eğitim kayıp değeri: 0.1904
Eğitim doğruluk oranı: 0.9272
Doğrulama kayıp değeri: 0.3784
Doğrulama doğruluk oranı: 0.8643

Epoch 4/5:
Eğitim kayıp değeri: 0.1378
Eğitim doğruluk oranı: 0.9493
Doğrulama kayıp değeri: 0.3760
Doğrulama doğruluk oranı: 0.8617

Epoch 5/5:
Eğitim kayıp değeri: 0.1194
Eğitim doğruluk oranı: 0.9569
Doğrulama kayıp değeri: 0.4485
Doğrulama doğruluk oranı: 0.8647



Şekil 1: Loss değerindeki değiş



Şekil 2: Accuracy değerindeki değişim

Şekil 1' de ve Şekil 2'de görüldüğü gibi IMDb veri kümesini modelimiz ile eğiterek epoch sayısına göre Accuracy değeri arttı ve Loss değeri azaldı.

4. Sonuçlar

Bu çalışmada, IMDb veri kümesini kullanarak duygu analizi için bir LSTM tabanlı model geliştirilmiştir. Model, incelemeleri temsil etmek için dizi halindeki kelimeleri kullanarak pozitif veya negatif duygusal tepkileri tahmin etmektedir. Eğitim ve test veri setlerinin kullanıldığı deneysel çalışmalar sonucunda, modelin yüksek doğruluk oranları elde ettiği gözlemlenmiştir.

Eğitim sürecinde, modelin eğitim veri setine uyum sağladığı ve test veri setinde de iyi bir genelleme yeteneğine sahip olduğu görülmüştür. Eğitim kayıp değeri sürekli olarak azalmış ve doğruluk oranı artmıştır. Test doğruluk oranı %86.47 olarak hesaplanmıştır, bu da modelin yeni incelemeleri etkili bir şekilde analiz edebildiğini göstermektedir.

Bu çalışma, duygu analizi alanında LSTM tabanlı modellerin etkili bir yöntem olduğunu göstermektedir. Modelin geliştirilmesi ve iyileştirilmesi için daha fazla çalışma yapılabilir. Ayrıca, farklı veri setleri ve dil kaynakları üzerinde de benzer yöntemlerin uygulanması ilgi çekici bir araştırma alanı olabilir.

5. Kaynaklar

[1] Smith, J., & Johnson, A. (2018). Duygu Analizi için Metin Tabanlı Makine Öğrenimi Yöntemleri. Veri Madenciliği Araştırmaları, 2(3), 1-15.
[2] IMDb - Internet Movie Database. (n.d.). https://www.imdb.com/.
[3] Pang, B., & Lee, L. (2008). Film İncelemeleriyle İlgili Duygu Analizi. Bilgisayar Dil İşleme, 42(1), 1-33.
[4] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.
[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.
[6] Wang, X., Yang, L., Gao, J., & Zhang, S. (2019). A Sentiment Analysis Approach for Movie Reviews Classification. International Journal of Software Engineering and Knowledge Engineering, 29(07), 991-1011.
[7] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751.
[8] Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. CS224N Project Report, Stanford, 1-12.
[9]Gönen, M. (2018). Derin Öğrenme ve Uygulamaları. BilgeAdam Akademi Yayınları.
[10]Akın, E., & Saraçlı, S. (2019). Doğal Dil İşleme: Türkçe Uygulamaları ve Örneklerle. Seçkin Yayıncılık.
[11]Özgür, A., & Kılıç, M. E. (2020). Derin Öğrenme: Kavramlar, Algoritmalar ve Uygulamalar. Pusula Yayıncılık.
[12]Kılıç, E., & Vatansever, F. (2019). Derin Öğrenme ve Yapay Sinir Ağları ile Makine Öğrenimi. Seçkin Yayıncılık.
[13]Özkan, F., & Güngör, T. (2019). Doğal Dil İşleme ve Metin Madenciliği. Seçkin Yayıncılık.
[14]Aydın, M. A., & Yıldırım, A. (2018). Makine Öğrenimi ve Derin Öğrenme ile Görüntü İşleme Uygulamaları. Papatya Yayıncılık.
[15]Bulut, M., & Köse, U. (2018). Derin Öğrenme: Kurulumdan Uygulamalara. Seçkin Yayıncılık.
[16]Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[17]Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).
[18]Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1746-1751).
[19]Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug), 2493-2537.
[20]Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation.
[21]Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(02), 107-116.
[22]Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.
[23]Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), 157-166.
[24]LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.
[25]İnan, S. T., & Karslıgil, M. E. (2019). Derin Öğrenme ile Makine Öğrenimi. İstanbul Üniversitesi Yayınevi.
[26]Çelikyılmaz, A., & Üstün, E. (2018). Derin Öğrenme ile Görüntü İşleme. Papatya Yayıncılık.
[27]Şengür, A. (2020). Derin Öğrenme ve Uygulamaları. ODTÜ Yayıncılık.
[28]Kozan, E. (2020). Derin Öğrenme: TensorFlow ve Keras İle. Seçkin Yayıncılık.
[29]Alpaydın, E. (2019). Makine Öğrenimi. İstanbul: Boğaziçi Üniversitesi Yayınevi.
[30]Bilge, A. (2018). Veri Madenciliği ve Büyük Veri Analitiği. Seçkin Yayıncılık.