MDPI

Article

Derin Öğrenme Modelleriyle Haber Sınıflandırılması

Aleyna KOCABEY [1], Burhan KORKMAZ [2], Ömer ÖZCAN [3], Hilmi GÜNER [4]

Abstract: Bu çalışmada, ileri makine öğrenimi tekniklerini kullanarak gazete haberlerini farklı kategorilere sınıflandırmayı araştırıyoruz. Veri setimiz, lemmatize text ve sınıf kategorilerine ayrılmış 1.5 milyon lemmatize edilmiş metinden oluşmaktadır.9 farklı sınıf bulunmaktadır bunlar; Siyaset, Magazin, Spor, Bilim Teknoloji, Finans Ekonomi, Kültür Sanat, Sağlık, Çevre, Turizm. Üç farklı modeli uyguluyor ve karşılaştırıyoruz: Evrişimli Sinir Ağları (CNN), Uzun Kısa Süreli Bellek Ağları (LSTM) ve Hafif Gradyan Artırma Makinesi (LightGBM).

Keywords: NLP; Deep Learning; Classification; Model

1. Introduction

Dijital çağın hızlı gelişimi, haber kaynaklarının sayısında büyük bir artışa yol açmış ve bilgiye erişimi kolaylaştırmıştır. Bu durum, haberlerin doğru bir şekilde sınıflandırılmasını ve ilgili kategorilere ayrılmasını önemli hale getirmiştir. Haberlerin türlerine göre sınıflandırılması, okuyucuların ilgilendikleri konuları daha hızlı bulmalarını sağlarken, aynı zamanda medya kuruluşlarının içerik yönetimini ve dağıtımını daha verimli hale getirir. Geleneksel yöntemlerle manuel olarak yapılan bu sınıflandırma işlemi, büyük veri setleri söz konusu olduğunda oldukça zaman alıcı ve hata yapmaya açık hale gelmektedir.

Bu çalışmada, gazete haberlerini otomatik olarak sınıflandırmak amacıyla makine öğrenimi ve derin öğrenme tekniklerini kullanıyoruz. Özellikle, Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM) [1] ve Light Gradient Boosting Machine (LightGBM) modellerinin performanslarını karşılaştırarak, hangi modelin bu görevde daha etkili olduğunu belirlemeyi hedefliyoruz[2]. CNN ve LSTM gibi derin öğrenme modelleri, metin verilerindeki karmaşık yapıları ve bağlamları öğrenme kapasitesiyle öne çıkarken, LightGBM modeli yüksek boyutlu verilerde hızlı ve etkili bir sınıflandırma sağlar.

Çalışmamızda kullanılan veri seti, 1.5 milyon lemmatize edilmiş gazete haberinden oluşmaktadır. Bu haberler, magazin, siyaset gibi çeşitli kategorilere ayrılmıştır. Veri ön işleme adımlarında metin temizleme, tokenizasyon, padding ve embedding işlemleri gerçekleştirilmiştir. Bu ön işleme adımları, metin verilerini sayısal verilere dönüştürerek modellerin eğitilmesi için uygun hale getirmektedir.

Bu çalışmanın amacı, gazete haberlerinin türlerini doğru ve hızlı bir şekilde sınıflandırmak için farklı modellerin etkinliğini değerlendirmek ve karşılaştırmaktır. Elde edilen sonuçlar, medya endüstrisi için daha verimli sınıflandırma yöntemleri geliştirilmesine katkıda bulunacaktır.

Citation: Derin Öğrenme Modelleriyle Haber Sınıflandırılması. *Journal Not Specified* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

2. Materials and Methods

2.0.1. Veri Seti

Bu çalışmada kullanılan veri seti, 1.5 milyon adet lemmatize edilmiş gazete haberinden oluşmaktadır. Haberler, "magazin" ve "siyaset" gibi çeşitli kategorilere ayrılmıştır. Her haber metni, haberin içeriğini temsil eden lemmatize kelimelerden oluşan bir dizidir. Veri seti, haberlerin türlerini doğru bir şekilde sınıflandırmak için yeterli çeşitliliği ve miktarı sağlamaktadır.

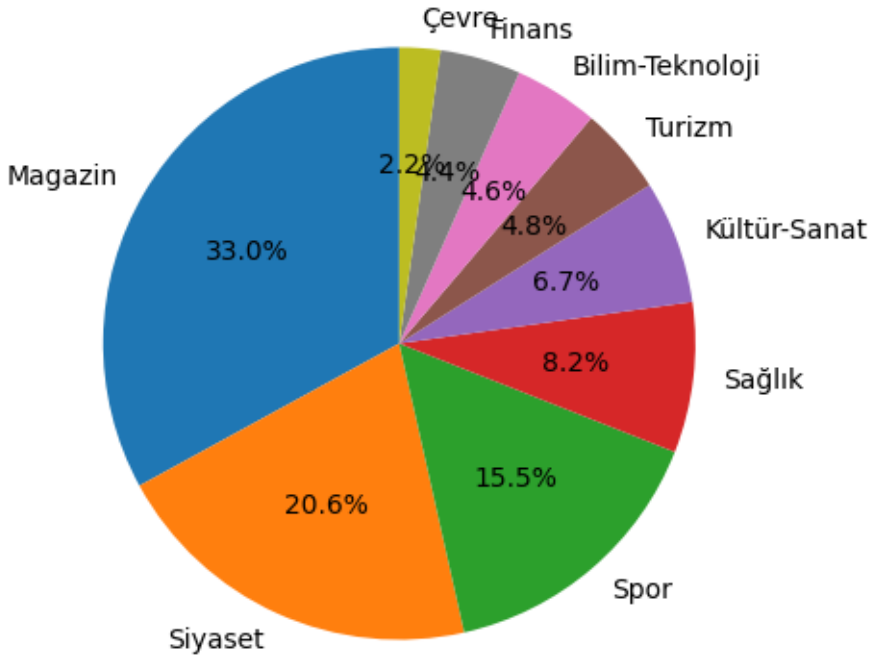

Figure 1. Veri setinin dağılımı

2.0.2. En sık tekrar eden kelimeler

Burada her sınıfın en çok tekrarlanan kelimelerini grafik halinde inceleyebiliriz. Dokuz sınıfın en çok tekrarlanan tokenları:

Bilim Sınıfı	**Çevre**	**Finans**
insan	alan	art
dünya	proje	faiz
teknoloji	çevre	seviye
geliş	ülke	dolar
sahip	orman	yatırım
sistem	kaynak	piyasa
bilgi	iklim	enflasyon
önemli	enerji	bitcoin
bilim	deniz	değer
alan	değer	bank
Kültür Sanat	**Magazin**	**Sağlık**
sanat	oyun	sağlık
film	dizi	hastalık
oyun	şarkı	dr
dünya	yaşa	tedavi
müzik	haber	hasta
kültür	hayat	gerek
insan	iste	hastane
tarih	önce	önemli
alan	medya	çocuk
zaman	film	kişi
Turizm	**Siyaset**	**Spor**
turizm	parti	maç
otel	seçim	takım
bin	bu	oyun
ülke	genel	lig
turist	erdoğan	futbol
milyon	chp	oyna
art	belediye	sezon
alan	ülke	gol
sektör	aday	kulüp
antalya	konuş	transfer

Table 1. Kategorilerdeki En Çok Tekrar Eden 10 Kelime

2.0.3. Veri Ön İşleme

Veri seti üzerinde çeşitli ön işleme adımları uygulanmıştır [3]:

1. Temizlik: Metinler, durak kelimelerden (stop words) ve özel karakterlerden ve veri setinde en sık tekrar eden %10, en az tekrar eden %10 kelimelerden arındırılmıştır.

2. Lemmatizasyon: Metinler, lemmatize edilerek her kelimenin kök hali elde edilmiştir.

3. Tokenizasyon: Metinler, kelime dizilerine dönüştürülmüştür.

4. Padding: Metinler, sabit bir uzunlukta olacak şekilde pad edilmiştir.

5. Embedding: Kelimeler, sayısal vektörler haline getirilmiştir. Pre-trained word embeddings (örn. Word2Vec, GloVe) veya rastgele başlatılmış embeddings kullanılmıştır.

2.0.4. Modeller

Bu çalışmada üç farklı model kullanılmıştır: CNN, LSTM ve LightGBM.

Convolutional Neural Networks (CNN) CNN, metin verilerini sınıflandırmak için kullanılan derin öğrenme modellerindendir. CNN modelinin yapısı şu şekildedir:

- Embedding Katmanı: Metin verilerini sayısal vektörlere dönüştürmek için kullanılır.
- Evrişim Katmanı (Convolutional Layer): Metin verilerindeki yerel örüntüleri öğrenir.
- Havuzlama Katmanı (Pooling Layer): Özellik haritalarını küçültür ve özetler.
- Flatten Katmanı: Özellik haritalarını tek boyutlu bir vektöre dönüştürür.
- Dense Katmanı: Tam bağlı katmanlarla sınıflandırma yapılır.
- Çıkış Katmanı (Output Layer): Softmax aktivasyon fonksiyonu kullanılarak sınıflandırma yapılır.

Long Short-Term Memory Networks (LSTM) LSTM, sıralı veriler üzerinde çalışan bir tür RNN'dir. LSTM modelinin yapısı şu şekildedir:

- Embedding Katmanı: Metin verilerini sayısal vektörlere dönüştürmek için kullanılır.
- LSTM Katmanı: Metin verilerindeki uzun dönemli bağımlılıkları öğrenir.
- Dense Katmanı: Tam bağlı katmanlarla sınıflandırma yapılır.
- Çıkış Katmanı (Output Layer): Softmax aktivasyon fonksiyonu kullanılarak sınıflandırma yapılır.

LightGBM LightGBM, gradient boosting framework'lerinden biridir ve özellikle büyük veri setlerinde ve yüksek boyutlu özelliklerde iyi performans gösterir. LightGBM modelinin yapısı şu şekildedir:

- Veri Dönüşümü: Metin verileri TF-IDF veya CountVectorizer kullanılarak sayısal verilere dönüştürülmüştür.
- Model Eğitimi: LightGBM, boosting algoritması kullanarak model eğitimi yapar.
- Parametre Ayarları: Modelin performansını artırmak için hiperparametre optimizasyonu yapılmıştır.

2.0.5. Eğitim ve Değerlendirme

Modellerin performansını değerlendirmek için veri seti eğitim, test ve doğrulama olarak üçe ayrılmıştır (%60 eğitim, %20 test, %20 valid). Modeller, eğitim seti üzerinde eğitilmiş ve test seti üzerinde değerlendirilmiştir. Değerlendirme metrikleri olarak doğruluk (accuracy), kesinlik (precision), geri çağırma (recall) ve F1 skoru kullanılmıştır.

2.0.6. Yazılım ve Donanım

Tüm modeller, Python programlama dili kullanılarak ve Keras, TensorFlow, scikit-learn gibi kütüphanelerle geliştirilmiştir. Eğitim ve değerlendirme işlemleri, Google Colab gibi yüksek RAM ve Hızlı GPU temin eden ortamda çalıştırılmıştır.

3. Conclusions

3.1. Streamlit

CNN modeli kullanarak hazırladığımız Streamlit sayfası.

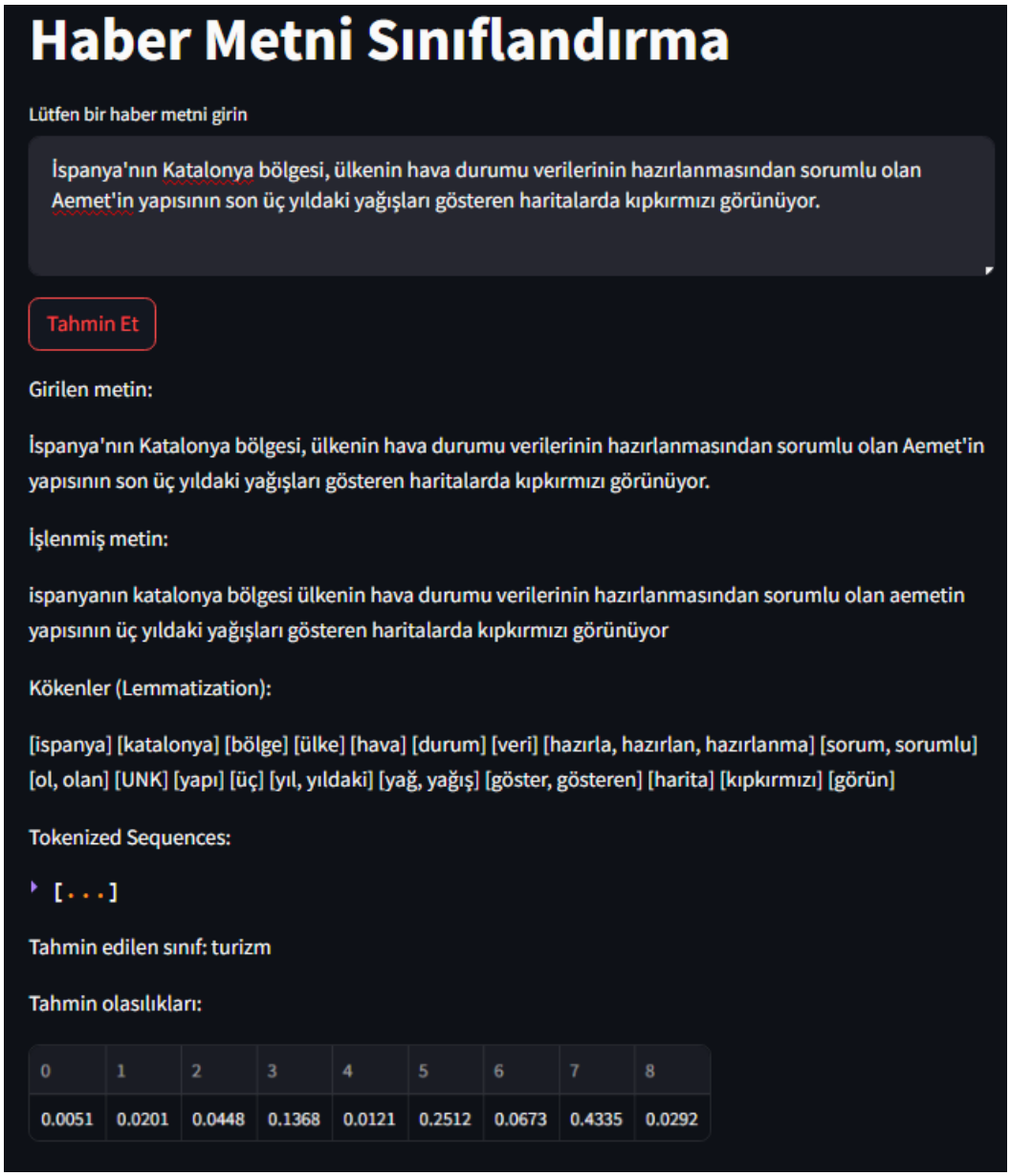

Figure 2. Streamlit sayfası

3.2. Model Sonuçları

Her üç modelin sonuçlarını gösteren tabloları inceleyebiliriz.

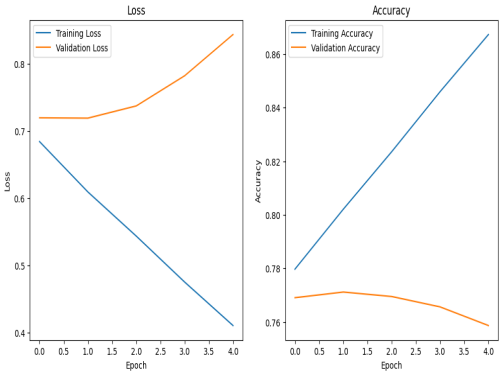

Figure 3. CNN modelinin doğruluk ve kayıp grafikleri

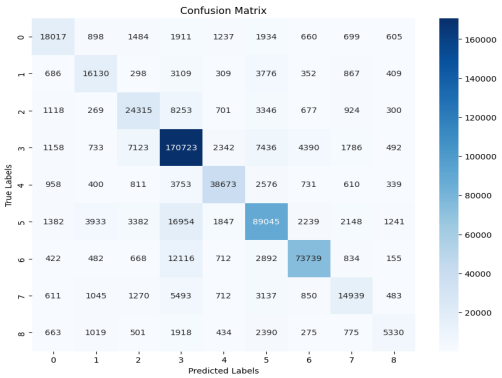

Figure 4. CNN modelinin Karmaşıklık Matrisi

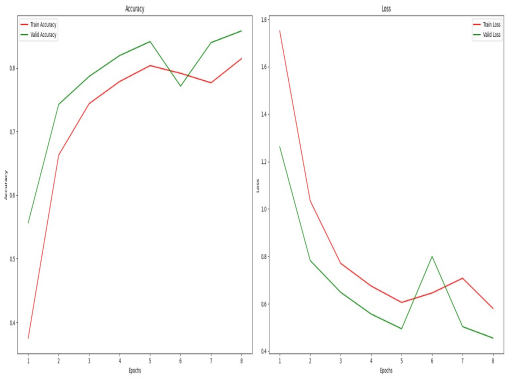

Figure 5. LSTM modelinin doğruluk ve kayıp grafikleri

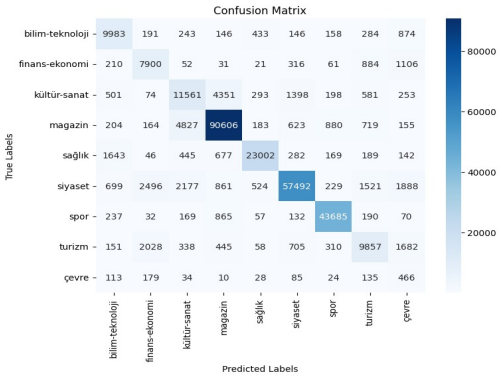

Figure 6. LSTM modelinin Karmaşıklık Matrisi

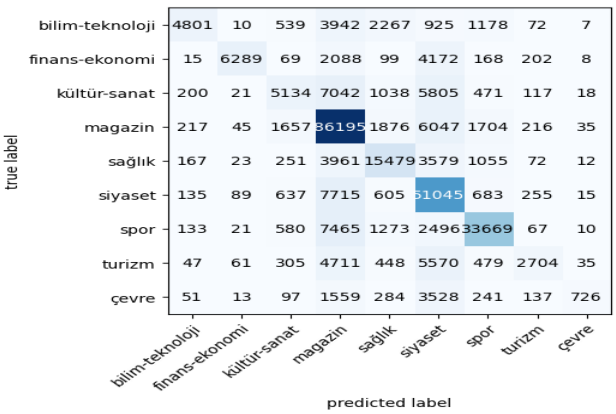

Figure 7. LigthGBM modelinin Karmaşıklık Matrisi

Model	Train Acc.	Test Acc.	Precision	Recall	F1 Score
CNN	0.8674	0.7587	0.71	0.67	0.75
LSTM	0.8147	0.86	0.85	0.86	0.85
LighGBM	0.70	0.69	0.71	0.69	0.67

Table 2. Model Değerleri

Bu çalışmada, gazete haberlerini türlerine göre otomatik olarak sınıflandırmak için üç farklı makine öğrenimi modelinin performansını karşılaştırdık: Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM) ve Light Gradient Boosting Machine (LightGBM) [2]. Her bir modelin farklı güçlü yönleri ve zayıf yönleri olduğu gözlemlenmiştir.

CNN ve LSTM modelleri, metin verilerindeki karmaşık örüntüleri ve uzun dönemli bağımlılıkları öğrenme kapasitesi sayesinde başarılı sonuçlar elde etmiştir [1]. CNN modeli, özellikle yerel özellikleri ve bağlamı yakalama konusunda etkili olmuş, LSTM modeli ise metinler arasındaki sıralı ve bağlamsal ilişkileri iyi bir şekilde modelleyebilmiştir. LightGBM modeli ise, yüksek boyutlu ve büyük veri setlerinde hızlı ve etkili bir sınıflandırma sağlayarak dikkat çekmiştir.

Değerlendirme sonuçlarına göre, her üç model de %70 %80 doğruluk oranlarına ulaşmış ve metin sınıflandırma görevinde başarılı olmuştur. Ancak, belirli veri türlerine ve sınıflandırma görevlerine göre her modelin performansı değişiklik göstermiştir.Veri Setinin Dengesiz oluşundan dolayı validation ve test sonuçlarında iyi sonuçlar gözlemlenememiştir.Overfitting durumuyla karşı karşıya kalınmıştır.

Bu çalışmanın sonuçları, gazete haberlerinin türlerine göre sınıflandırılması için makine öğrenimi ve derin öğrenme modellerinin etkinliğini göstermektedir. CNN ve LSTM modelleri, derin öğrenme mimarileri sayesinde daha yüksek doğruluk ve F1 skorlarına ulaşırken, LightGBM modeli hız ve verimlilik açısından üstünlük sağlamıştır. Bu sonuçlar, medya endüstrisi için daha verimli ve doğru sınıflandırma sistemleri geliştirilmesine katkıda bulunabilir.

Gelecekteki çalışmalar, bu modellerin daha büyük ve çeşitlendirilmiş veri setlerinde test edilmesi ve hiperparametre optimizasyonu ve model mimarisi üzerinde daha fazla çalışma yapılması yönünde ilerleyebilir. Ayrıca, model performansını artırmak için hibrit modellerin ve transfer öğrenme yöntemlerinin kullanılması da araştırılabilir.

References

1. Li, C.; Zhan, G.; Li, Z. News text classification based on improved Bi-LSTM-CNN. In Proceedings of the 2018 9th International conference on information technology in medicine and education (ITME). IEEE, 2018, pp. 890–893.
2. Ibrahim, Y.; Okafor, E.; Yahaya, B.; Yusuf, S.M.; Abubakar, Z.M.; Bagaye, U.Y. Comparative study of ensemble learning techniques for text classification. In Proceedings of the 2021 1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS). IEEE, 2021, pp. 1–5.
3. Vijayarani, S.; Ilamathi, M.J.; Nithya, M.; et al. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks* **2015**, *5*, 7–16.