

Kitap Türü Tahmini İçin Metin Sınıflandırma Raporu

1. Giriş

Bu çalışmada, kitap açıklamalarını kullanarak kitap türlerini tahmin eden bir metin sınıflandırma modeli geliştirilmiştir. Çalışma kapsamında farklı veri ön işleme teknikleri, vektörleştirme yöntemleri ve derin öğrenme modelleri kullanılarak en iyi performansı sağlayan yaklaşım belirlenmiştir.

1.1. Veri Ön İşleme

Veri seti üzerinde aşağıdaki ön işleme adımları gerçekleştirilmiştir:

- **Metinleri Küçük Harfe Dönüştürme:** Bütün metinler küçük harfe çevrilmiştir.
- **Noktalama İşaretlerini Kaldırma:** Noktalama işaretleri temizlenmiştir.
- **Boşlukları, URL'leri, Özel Karakterleri kaldırma:** web siteleri linkleri temizleşmiştir.
- **Stop-word Temizliği:** İngilizce için etkisiz kelimeler kaldırılmıştır.
- **Stemming/Lemmatization:** Kelimeler köklerine indirgenmiştir.
- **Veri Görselleştirme:** Kelime sıklıkları, kategori dağılımları ve kelime bulutları analiz edilmiştir.

1.2. Vektörleştirme Yöntemleri

Aşağıdaki yöntemler kullanılarak metinler sayısal formata dönüştürülmüştür:

- **Bag of Words (BoW)**
- **TF-IDF (Term Frequency-Inverse Document Frequency)**
- **Word2Vec**
- **BERT (Bidirectional Encoder Representations from Transformers)**

1.3. Model Geliştirme ve Değerlendirme

Farklı vektörleştirme yöntemleriyle aşağıdaki derin öğrenme modelleri eğitilmiştir:

- **LSTM (Long Short-Term Memory)**
- **Multilayer Perceptron (MLP)**

Her model için hiperparametre optimizasyonu gerçekleştirilmiş, eğitim süreci izlenmiş ve değerlendirme metrikleri hesaplanmıştır.

1.4. Sonular ve Karşılařtırma

Her model ve vektörleřtirme yöntemi için ařağıdaki metrikler incelenmiřtir:

- Doğruluk (Accuracy)
- Kesinlik (Precision)
- Duyarlılık (Recall)
- F1-Skoru
- Karmaşıklık Matrisi
- Eğitim ve doğrulama kayıpları/grafikleri

2. Vektörleřtirme Yöntemleri ve Eğitim

2.1 Bag of Words (BoW)

2.1.1 Veri ön İşleme

Bag of words yöntemi için ařağıdaki ön işleme teknikleri kullanıldığında daha iyi sonuç verdiğı gözlenmiřtir. Veri arttırma yöntemi ile model daha da iyi performans vermiřtir.

- Metinleri Küçük Harfe Dönüřtürme
- Noktalama İşaretlerini Kaldırma
- Boşlukları, URL'leri, Özel Karakterleri kaldırma
- Stop-word Temizliğı
- Stemming/Lemmatization
- Veri arttırma

Veriler ařağıdaki veri arttırma yöntemleriyle çoğaltılmıřtır.

Eřanlamlı Değıřtirme (Synonym Replacement): Bu yöntem, metin içindeki kelimeleri eřanamlılarıyla değıřtirerek veri setini zenginleřtirir. Modelin, farklı kelime kullanımlarına karşı daha dayanıklı hale gelmesini ve genelleme yeteneğinin artmasını sağlar.

Rastgele Silme (Random Deletion): Bu yöntem, metin içindeki kelimeleri rastgele bir olasılıkla silerek veri setini çeřitlendirir. Modelin, eksik veya gürültülü verilere karşı daha sağlam hale gelmesini ve genelleme yeteneğinin artmasını sağlar.

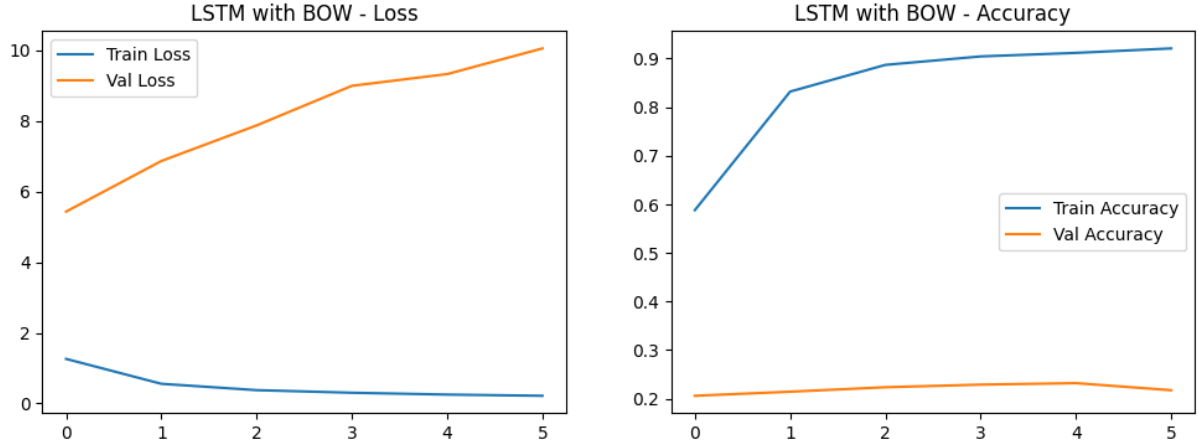
2.1.2 Model Geliřtirme ve Değerlendirme

Bag of Words yöntemi için çeřitli yöntemler ve optimize edilmiř modeller denenmiřtir. Optimizasyon yöntemlerinde model mimarisi değıřtirilmiř ve Regularization, Batch normalizasyon gibi yöntemler denenmiřtir. Ařağıda bu yöntemleri ve çıktıları yer almaktadır.

Ön işlenmiş veri + LSTM:

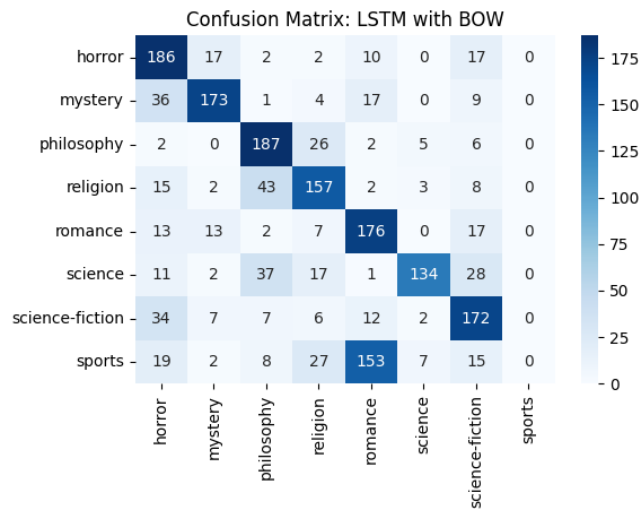
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
LSTM + BOW	0.6368	0.5850	0.6368	0.5987

Grafikler:



Grafikten de anlaşılacağı üzere aşırı öğrenme durumu söz konusu bu durumu çözmek için model mimarisini ya da ön işleme yöntemlerini değiştirmek gerekir.

Karmaşıklık Matrisi:

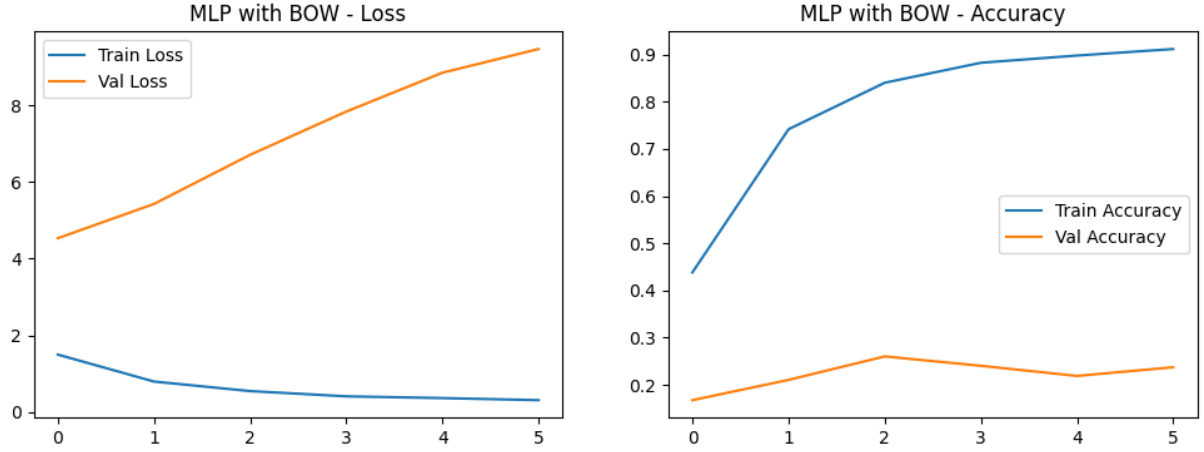


Model özellikle spor ve romantizm sınıflarını karıştırıyor.

Ön işlenmiş veri + MLP:

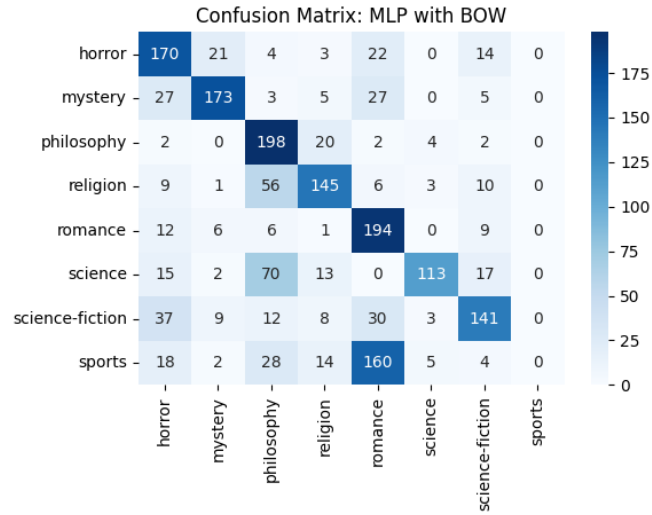
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
MLP + BOW	0.6093	0.5811	0.6093	0.5731

Grafikler:



aşırı öğrenme durumu söz konusu bu durumu çözmek için model mimarisini ya da ön işleme yöntemlerini değiştirmek gerekir.

Karmaşıklık Matrisi:

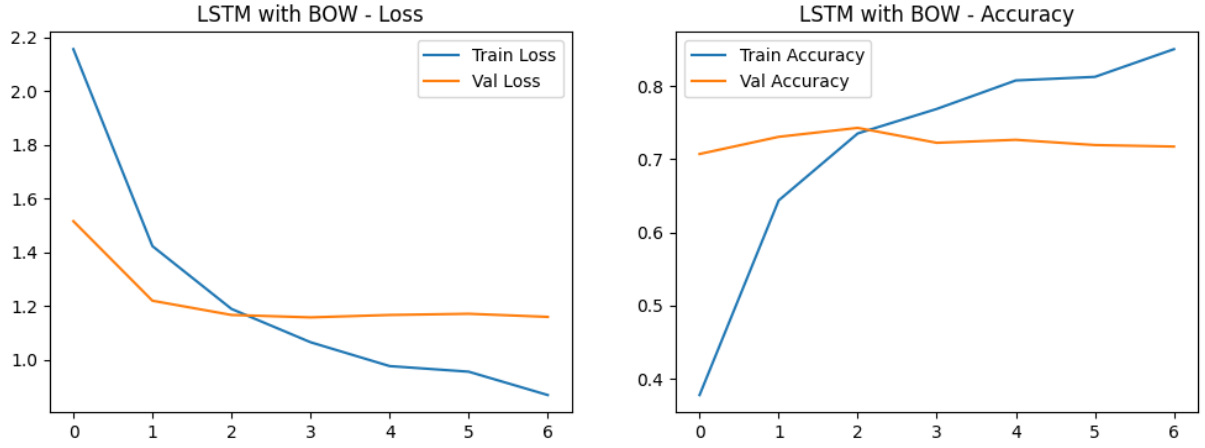


Model bu mimaride yine romantizm ve spor sınıflarını karıştırdı ek olarak psikoloji ve bilim sınıflarında da karışıklık söz konusu

Ön İşleme + Optimize edilmiş LSTM

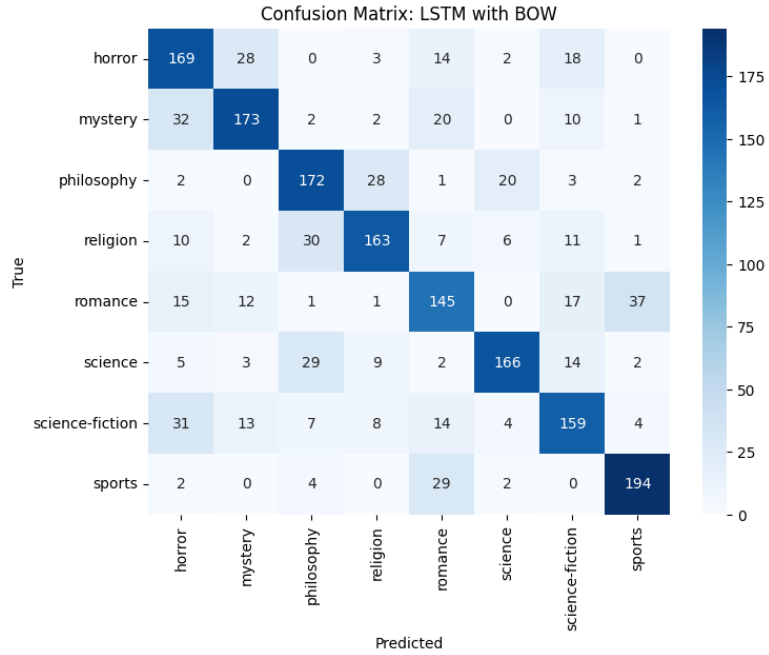
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(LSTM) + BOW	0.7319	0.7332	0.7319	0.7314

Grafikler:



Optimize edilmiş model ile performans arttı ve overfittingi önlemek için çeşitli yöntemlerin(Dropout, Regularization, LR lüçültmek) yanı sıra Erken durdurma ile modelin aşırı öğrenmesi azaltılmaya çalışıldı

Karmaşıklık Matrisi:

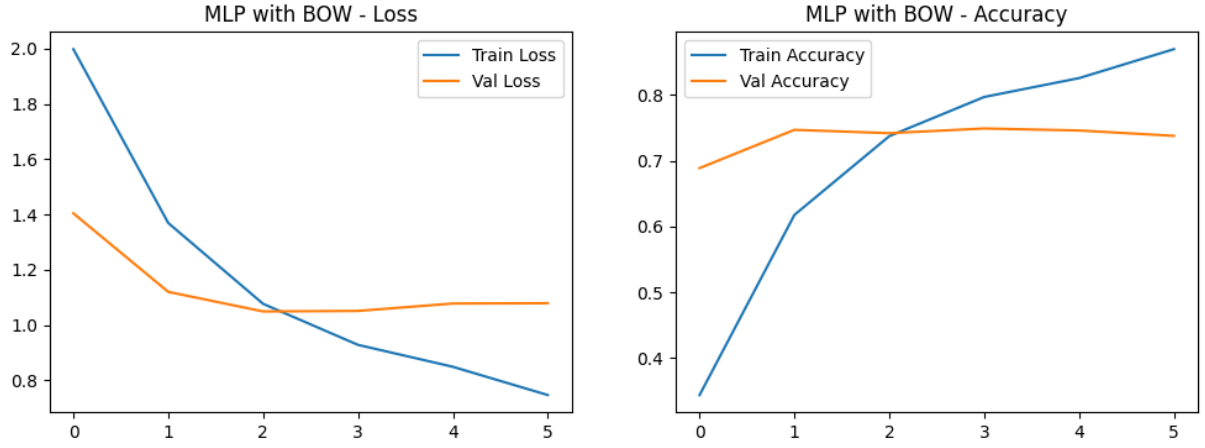


Optimize edilmiş model daha tutarlı bir sınıflandırma yapıyor. Spor ve romantizm karışıklığı önemli ölçüde düzelmiş durumda. Yine de spor-romance, science fiction-horror, horror-mystery gibi sınıflar karıştırılmakta

Ön İşleme + Optimize edilmiş MLP

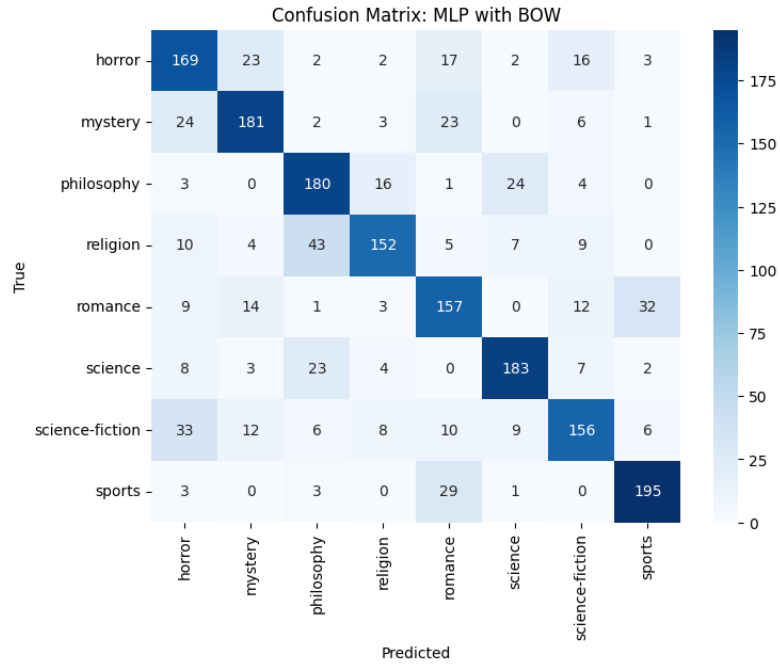
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(MLP) + BOW	0.7324	0.7339	0.7324	0.7323

Grafikler:



MLP modeli de LSTM gibi, model yeterince öğrenemiyor ve aşırı öğrenmeye kayıyor.

Karmaşıklık Matrisi:



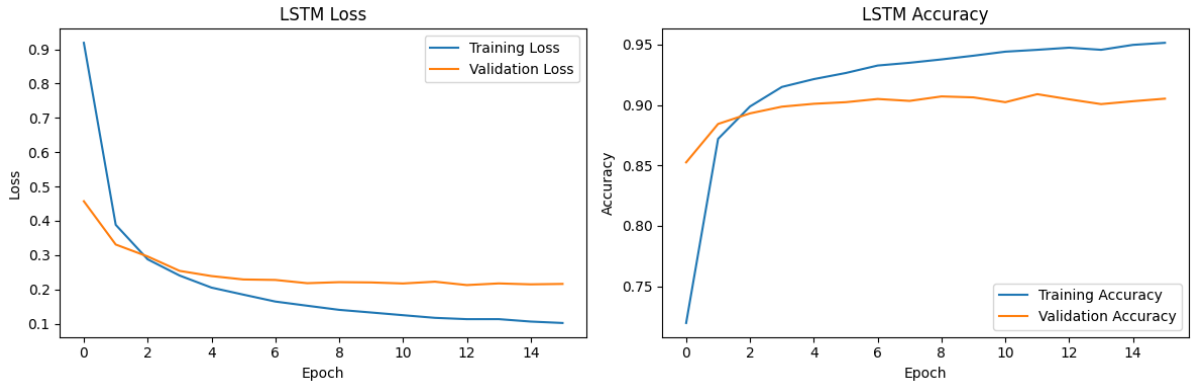
Optimize edilmiş model daha tutarlı bir sınıflandırma yapıyor. Spor ve romantizm karışıklığı önemli ölçüde düzelmiş durumda. Ama halen daha karışıklıklar mevcut. Horror ve science-fiction, philosophy ve religion gibi.

Ön işleme + Veri arttırma + Optimize LSTM:

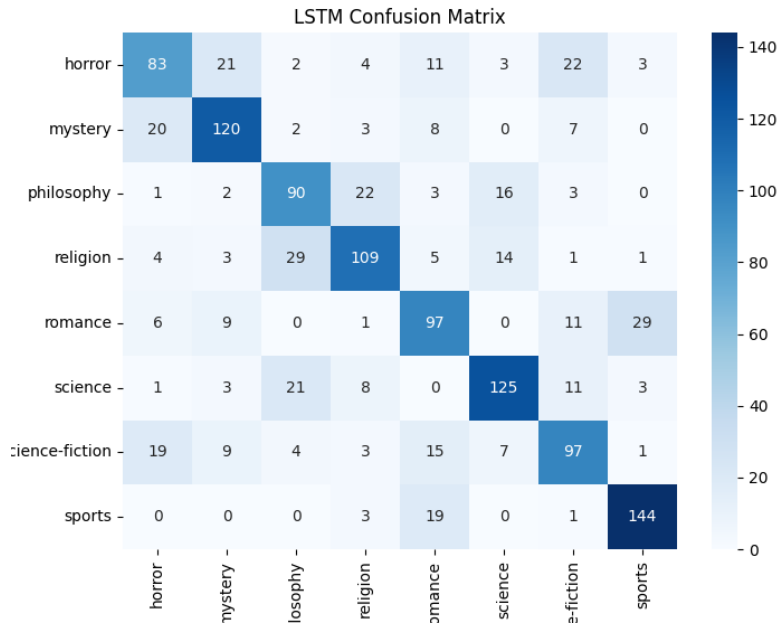
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(LSTM) + Veri arttırma +BOW	0.6871	0.6824	0.6837	0.6823

Veri arttırma yöntemi modelin performansında bir etki yaratmadı.

Grafikler:



Karmaşıklık Matrisi:



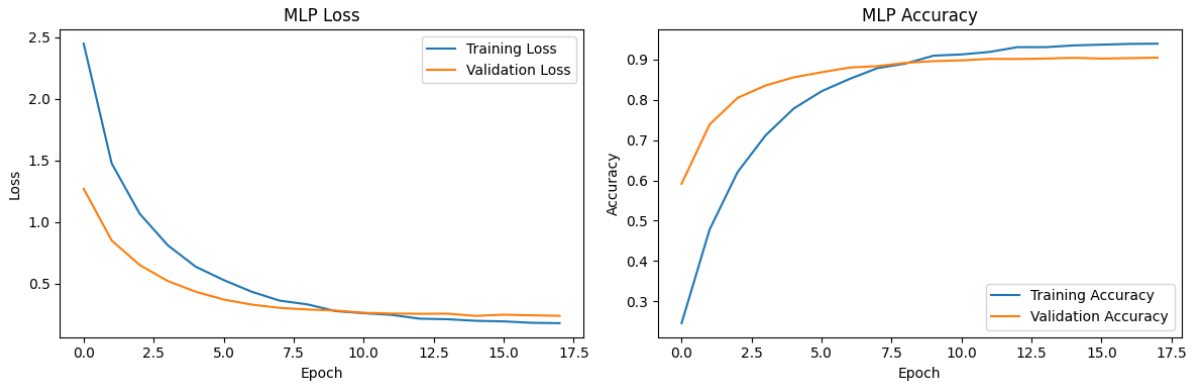
Sports-romance, horror-science iction gibi sınıflar halen daha diğer sınıflara göre daha fazla yanlışla sahip.

Ön işleme + Veri arttırma + Optimize MLP:

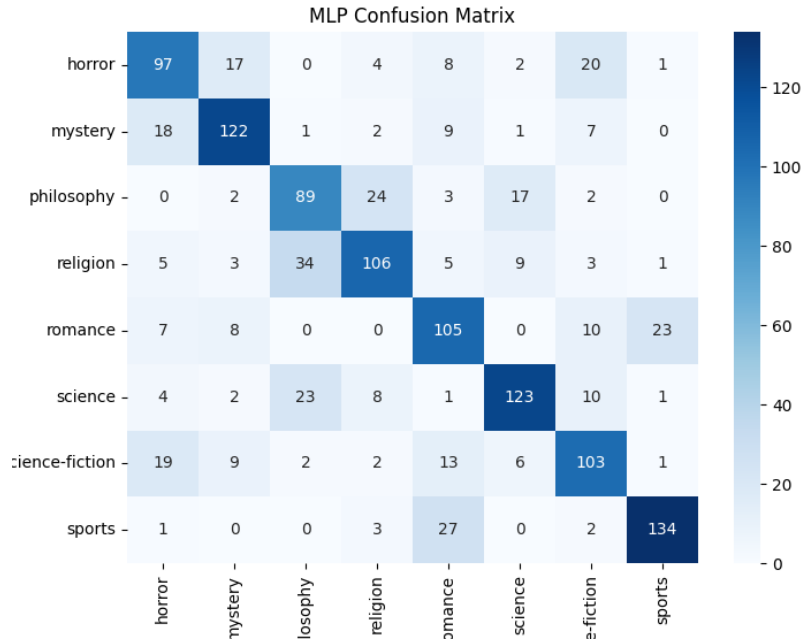
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(MLP) + Veri arttırma +BOW	0.6982	0.6976	0.6961	0.6962

Veri arttırma yöntemi modelin performansında bir etki yaratmadı

Grafikler:



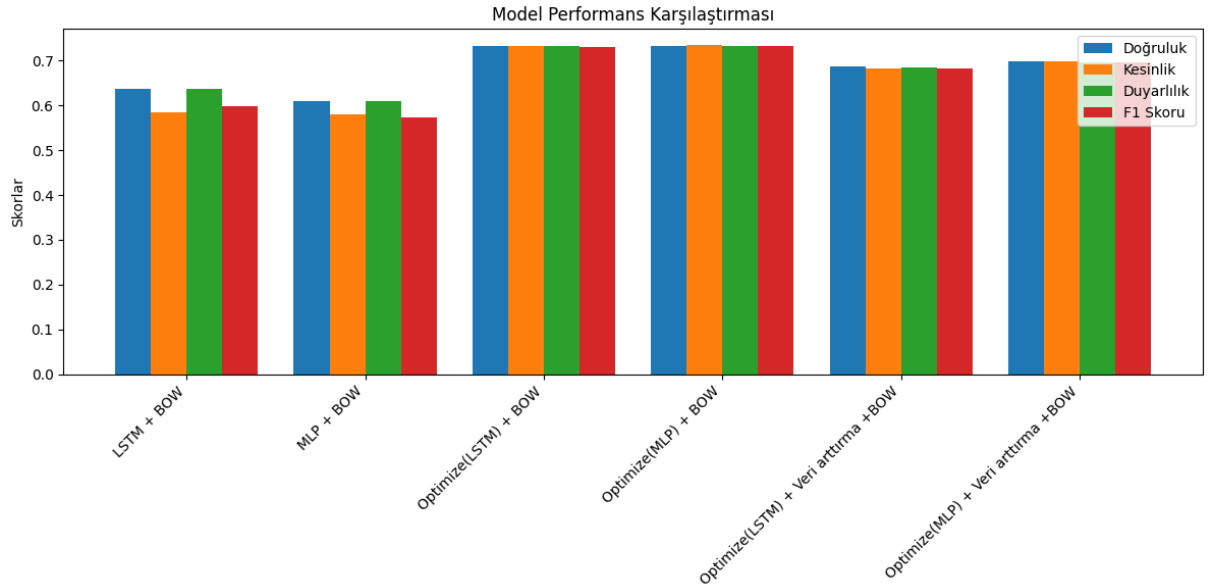
Karmaşıklık Matrisi:



LSTM modelinden farklı olarak religion-philosophy daha fazla karıştırılmakta. Bunun yanı sıra romance-sports 2. En çok karıştırılan

BoW için Nihai Tablo:

Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
LSTM + BOW	0.6368	0.5850	0.6368	0.5987
MLP + BOW	0.6093	0.5811	0.6093	0.5731
Optimize(LSTM) + BOW	0.7319	0.7332	0.7319	0.7314
Optimize(MLP) + BOW	0.7324	0.7339	0.7324	0.7323
Optimize(LSTM) + Veri arttırma +BOW	0.6871	0.6824	0.6837	0.6823
Optimize(MLP) + Veri arttırma +BOW	0.6982	0.6976	0.6961	0.6962



Genel olarak, optimize edilmiş modellerin (Optimize(LSTM) ve Optimize(MLP)) diğerlerine göre daha yüksek doğruluk, kesinlik, duyarlılık ve F1 skoruna sahip olduğu görülmektedir. Veri artırma tekniklerinin ise optimize edilmiş modellerin performansını bir miktar düşürdüğü gözlemlenmiştir.

2.2 TF-IDF

2.2.1 Veri ön İşleme

Tf-Idf yöntemi için aşağıdaki ön işleme teknikleri kullanıldığında daha iyi sonuç verdiği gözlenmiştir. Veri arttırma yöntemi ile model daha da iyi performans vermiştir.

- Metinleri Küçük Harfe Dönüştürme
- Noktalama İşaretlerini Kaldırma
- Boşlukları, URL'leri, Özel Karakterleri kaldırma
- Stop-word Temizliği
- Stemming/Lemmatization
- Veri arttırma

Veriler aşağıdaki veri arttırma yöntemleriyle çoğaltılmıştır.

Eşanlamlı Değiştirme (Synonym Replacement): Bu yöntem, metin içindeki kelimeleri eşanlamlılarıyla değiştirerek veri setini zenginleştirir. Modelin, farklı kelime kullanımlarına karşı daha dayanıklı hale gelmesini ve genelleme yeteneğinin artmasını sağlar.

Rastgele Silme (Random Deletion): Bu yöntem, metin içindeki kelimeleri rastgele bir olasılıkla silerek veri setini çeşitlendirir. Modelin, eksik veya gürültülü verilere karşı daha sağlam hale gelmesini ve genelleme yeteneğinin artmasını sağlar.

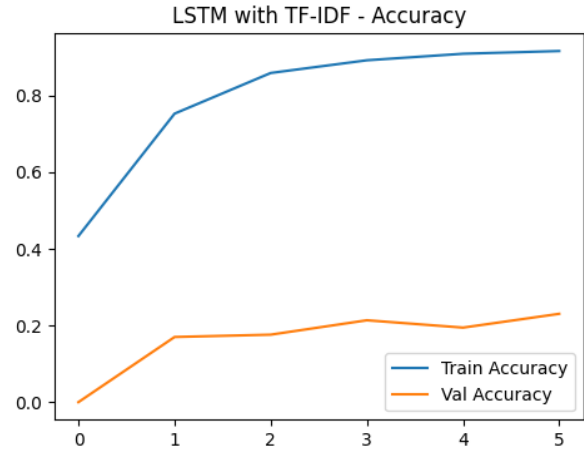
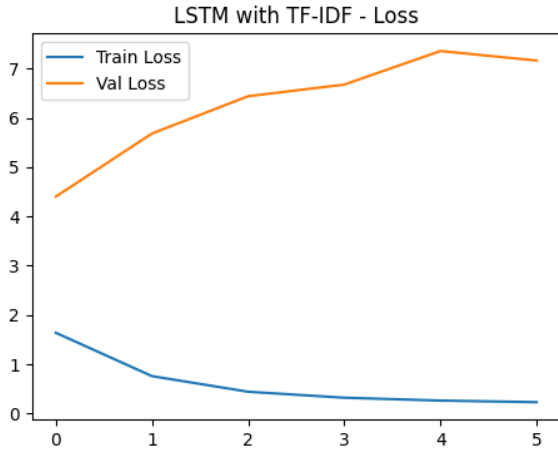
2.2.2 Model Geliştirme ve Değerlendirme

Tf-Idf yöntemi için çeşitli yöntemler ve optimize edilmiş modeller denenmiştir. Optimizasyon yöntemlerinde model mimarisi değiştirilmiş ve Regularization, Batch normalizasyon gibi yöntemler denenmiştir. Aşağıda bu yöntemleri ve çıktıları yer almaktadır.

Ön işleme + LSTM:

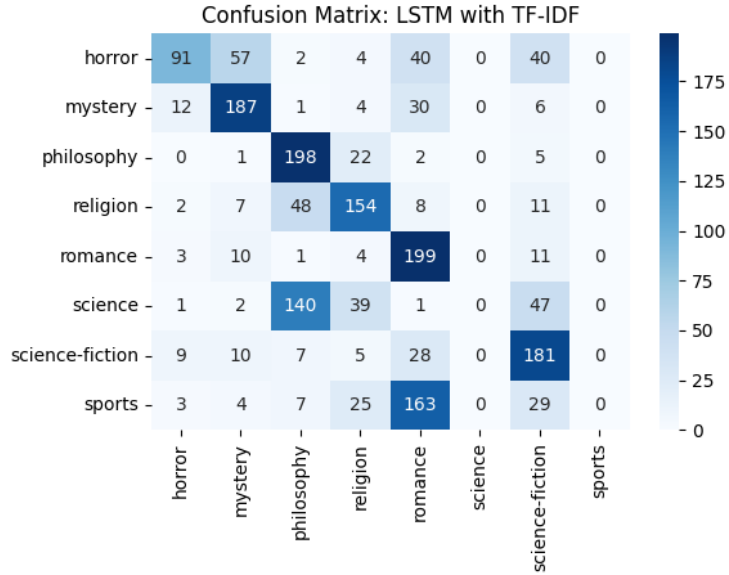
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
LSTM + TF-IDF	0.5427	0.4379	0.5427	0.4642

Grafikler:



Model performansı çok düşük ve aşırı öğrenme var

Karmaşıklık Matrisi:

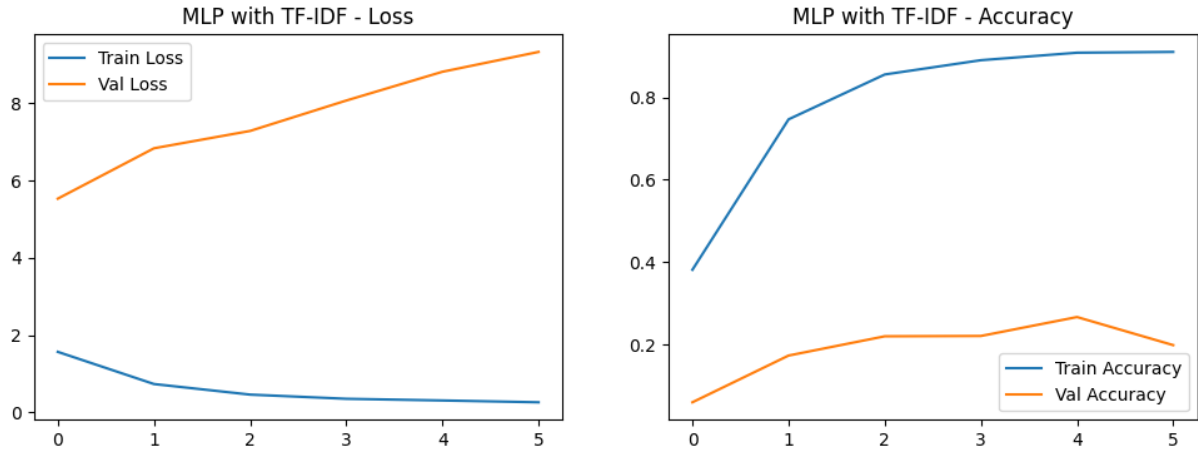


Model spor- romantizm, science-philosophy gibi sınıfları büyük ölçüde karıştırmaktadır. Science ve sports sınıfını hiç doğru tahmin edememiştir.

Ön işleme + MLP:

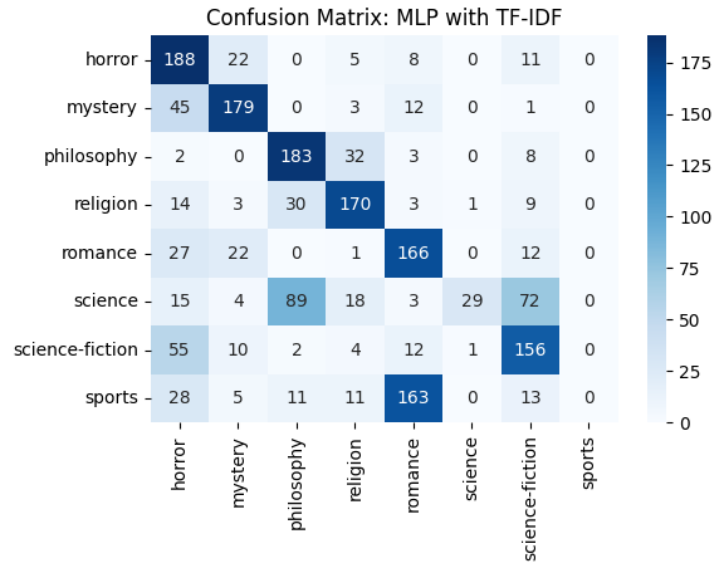
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
MLP + TF-IDF	0.5755	0.5556	0.5755	0.5167

Grafikler:



Model performansı çok düşük ve aşırı öğrenme var

Karmaşıklık Matrisi:

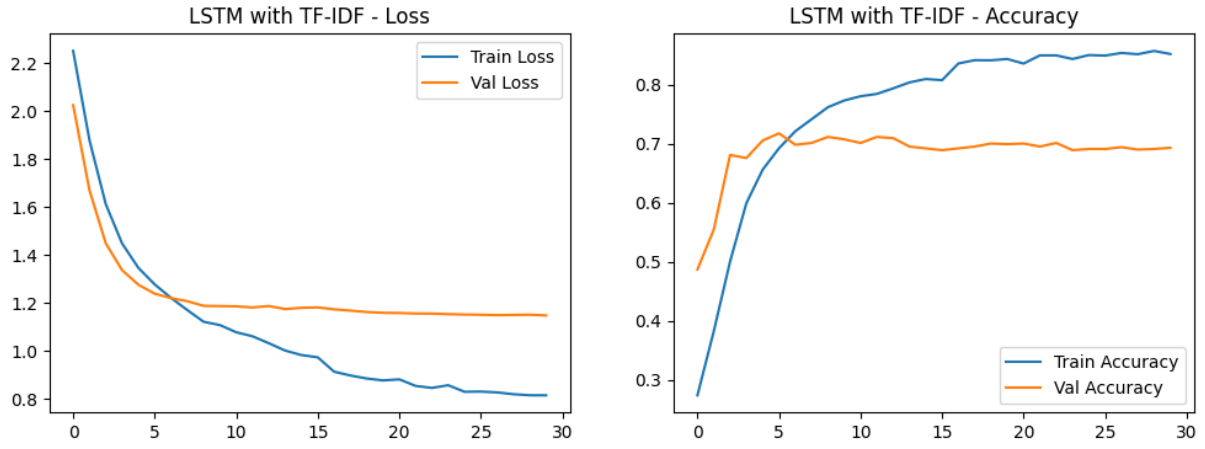


Model spor- romantizm, science-philosophy, horror-science fiction gibi sınıfları büyük ölçüde karıştırmaktadır. Science sınıfını neredeyse hiç doğru tahmin edememiştir. Sports sınıfını hiç doğru tahmin edememiştir.

Ön işleme + Optimize edilmiş LSTM:

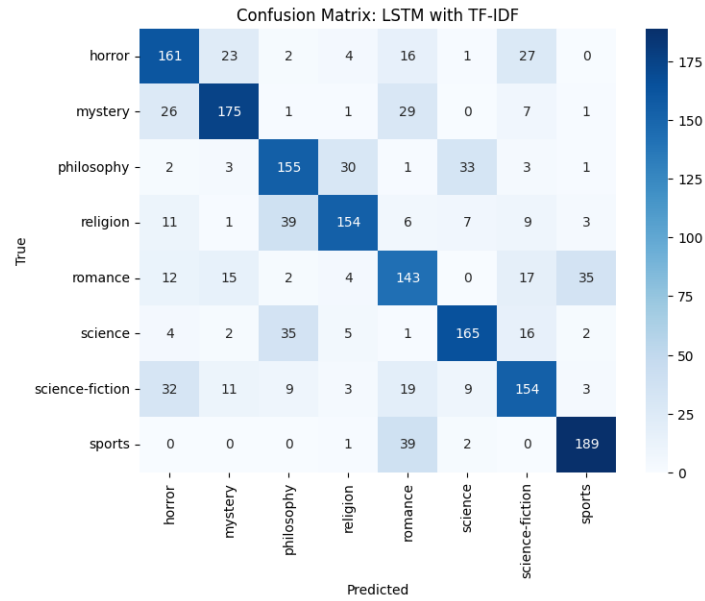
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(LSTM) + TF-IDF	0.7012	0.7037	0.7012	0.7016

Grafikler:



Model optimize edildikten sonra performansı artmıştır. Yine aşırı öğrenme söz konusu.

Karmaşıklık Matrisi:

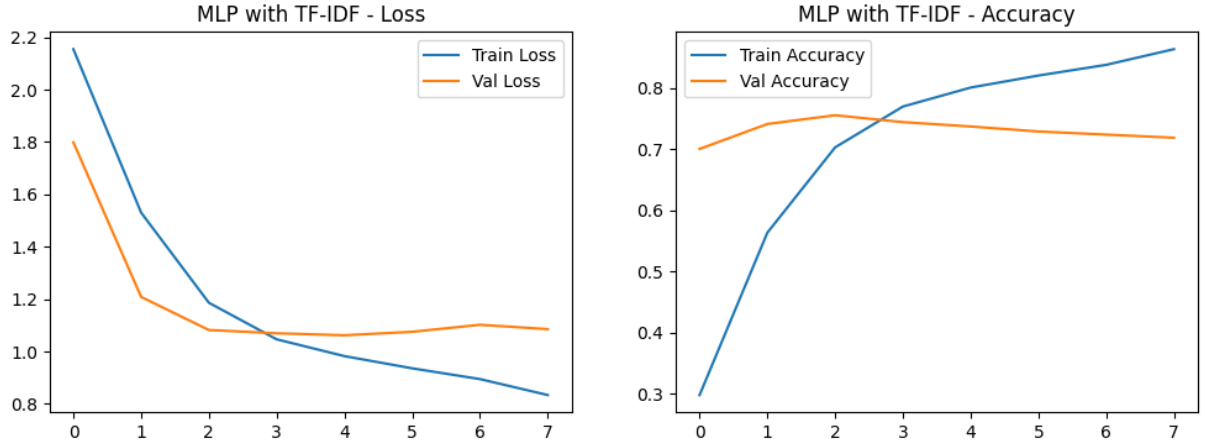


model optimize edildikten sonra tahmin edemediği sınıfları tahmin edebilmeye başlamıştır. Bunun yanı sıra sports-romance, religion-philosophy, horror-sciencefiction gibi sınıfları karıştırmaktadır.

Ön işleme + Optimize edilmiş MLP:

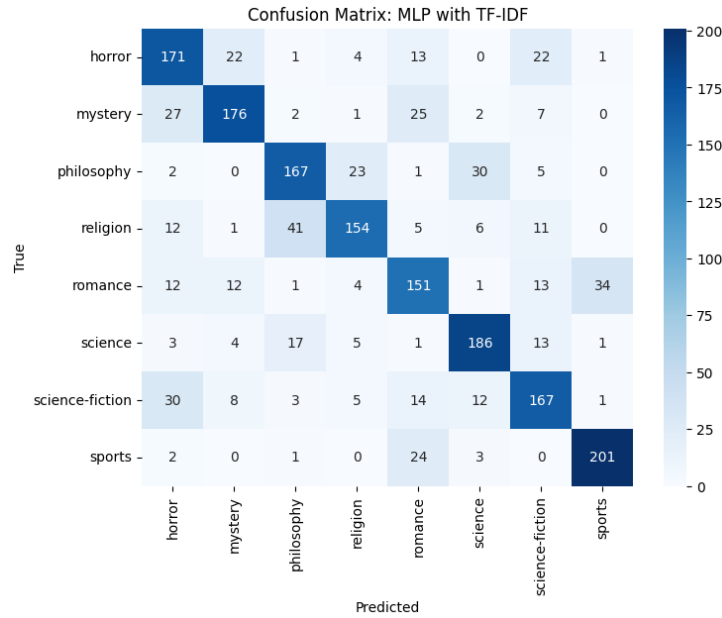
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(MLP) + TF-IDF	0.7378	0.7399	0.7378	0.7378

Grafikler:



Model optimize edildikten sonra performansı artmıştır. Ama yine aşırı öğrenme söz konusu.

Karmaşıklık Matrisi:



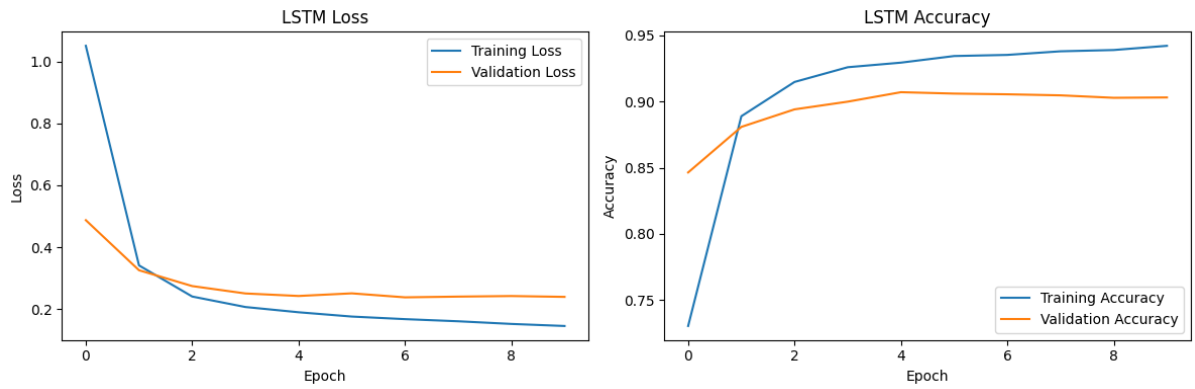
model optimize edildikten sonra tahmin edemediği sınıfları tahmin edebilmeye başlamıştır. Ama religion-philosophy sınıfı başta olmak üzere sports-romance, sciencefiction-horror gibi sınıfları karıştırmaktadır

Ön işleme + Veri arttırma + Optimize LSTM:

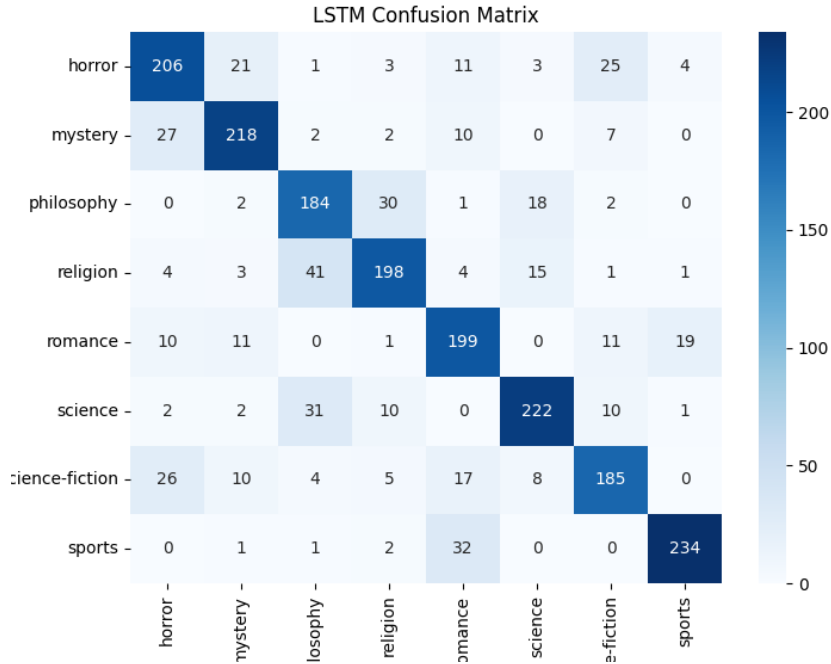
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(LSTM)+ Veri Arttırma + TF-IDF	0.7846	0.7850	0.7841	0.7839

Veri arttırma modelin performansını arttırmıştır

Grafikler:



Karmaşıklık Matrisi:



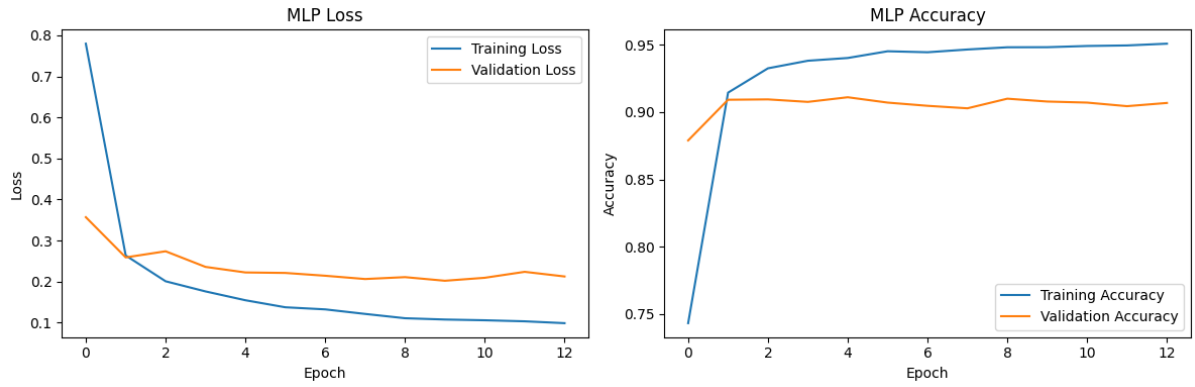
Veri arttırma modelin genelleme yeteneğini arttırsa da benzer sınıflar halen daha karıştırılmakta. Sports-romance, religion-philosophy gibi.

Ön işleme + Veri arttırma + Optimize MLP:

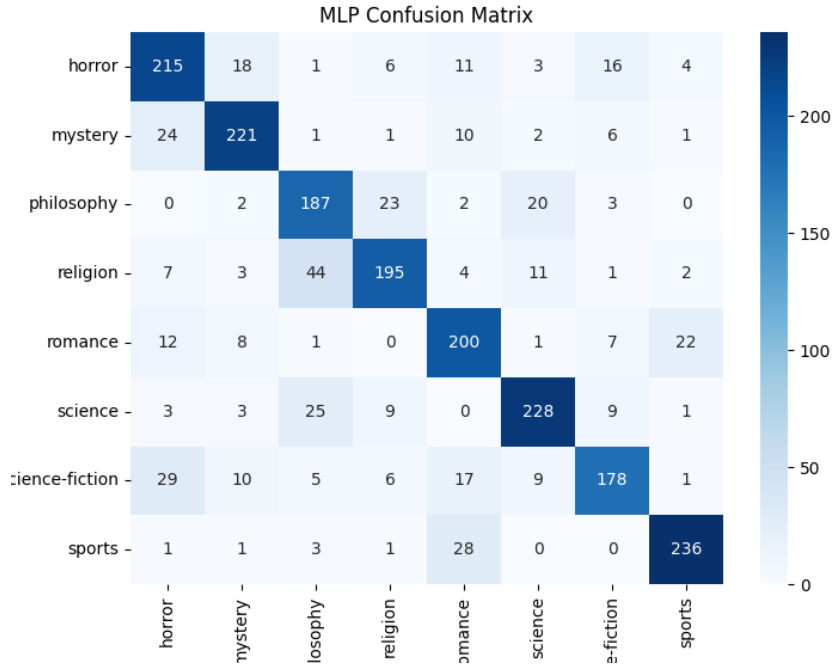
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(MLP)+ Veri Arttırma + TF-IDF	0.7912	0.7924	0.7905	0.7901

Veri arttırma sonrası modelin performansı artmıştır.

Grafikler:



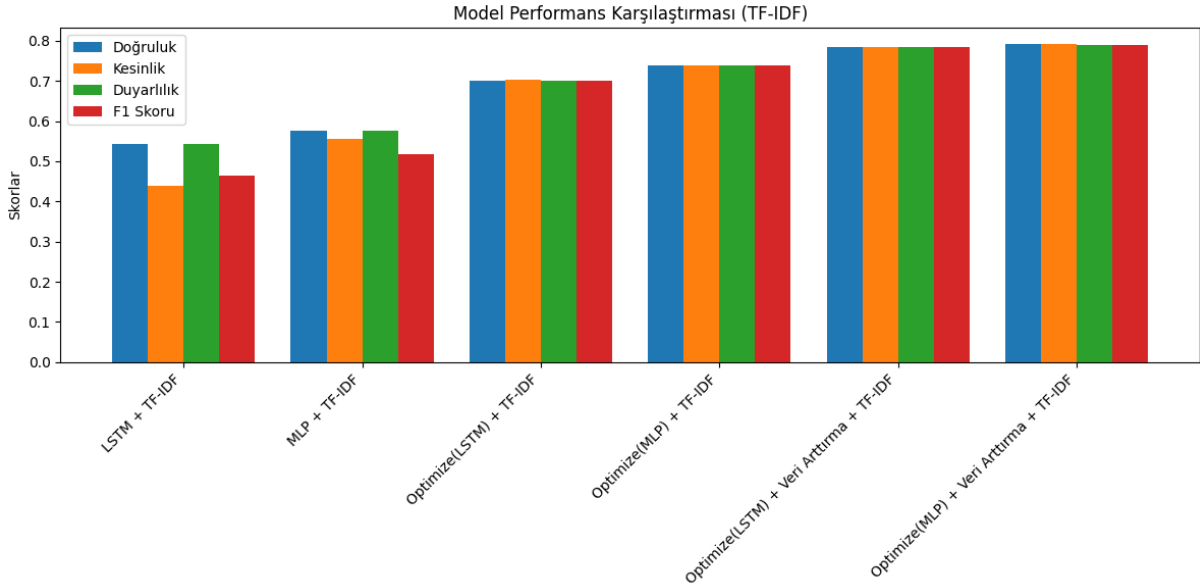
Karmaşıklık Matrisi:



Veri arttırma modelin genelleme yeteneğini arttırsa da benzer sınıflar halen daha karıştırılmakta. religion-philosophy gibi.

TF-IDF için Nihai Tablo:

Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
LSTM + TF-IDF	0.5427	0.4379	0.5427	0.4642
MLP + TF-IDF	0.5755	0.5556	0.5755	0.5167
Optimize(LSTM) + TF-IDF	0.7012	0.7037	0.7012	0.7016
Optimize(MLP) + TF-IDF	0.7378	0.7399	0.7378	0.7378
Optimize(LSTM)+ Veri Arttırma + TF-IDF	0.7846	0.7850	0.7841	0.7839
Optimize(MLP)+ Veri Arttırma + TF-IDF	0.7912	0.7924	0.7905	0.7901



Genel olarak, optimize edilmiş ve veri artırma teknikleri uygulanan modellerin diğerlerine göre daha yüksek doğruluk, kesinlik, duyarlılık ve F1 skoruna sahip olduğu görülmektedir. Veri artırma tekniklerinin, optimize edilmiş modellerin performansını önemli ölçüde artırdığı gözlemlenmiştir.

2.3 Word2Vec

2.3.1 Veri ön işleme

Word2Vec yöntemi için Bow ve TF-IDF yöntemlerinden farklı ön işleme yöntemleri uygulanmıştır. Word2Vec bağlamsal anlamı yakalayabildiği için stop words'ler çıkarılmamıştır. Aynı şekilde kelimeler köklerine indirgenmemiştir. Noktalama işaretleri kaldırılmamıştır. Sadece aşağıdaki yöntemler uygulandığında modelin daha iyi performans verdiği gözlemlenmiştir. Veri arttırma yöntemi ile model daha da iyi performans vermiştir.

- Metinleri Küçük Harfe Dönüştürme
- Boşlukları, URL'leri, Özel Karakterleri kaldırma
- Veri arttırma

Veriler aşağıdaki veri arttırma yöntemleriyle çoğaltılmıştır.

Eşanlamlı Değiştirme (Synonym Replacement): Bu yöntem, metin içindeki kelimeleri eşanlamlılarıyla değiştirerek veri setini zenginleştirir. Modelin, farklı kelime kullanımlarına karşı daha dayanıklı hale gelmesini ve genelleme yeteneğinin artmasını sağlar.

Rastgele Silme (Random Deletion): Bu yöntem, metin içindeki kelimeleri rastgele bir olasılıkla silerek veri setini çeşitlendirir. Modelin, eksik veya gürültülü verilere karşı daha sağlam hale gelmesini ve genelleme yeteneğinin artmasını sağlar.

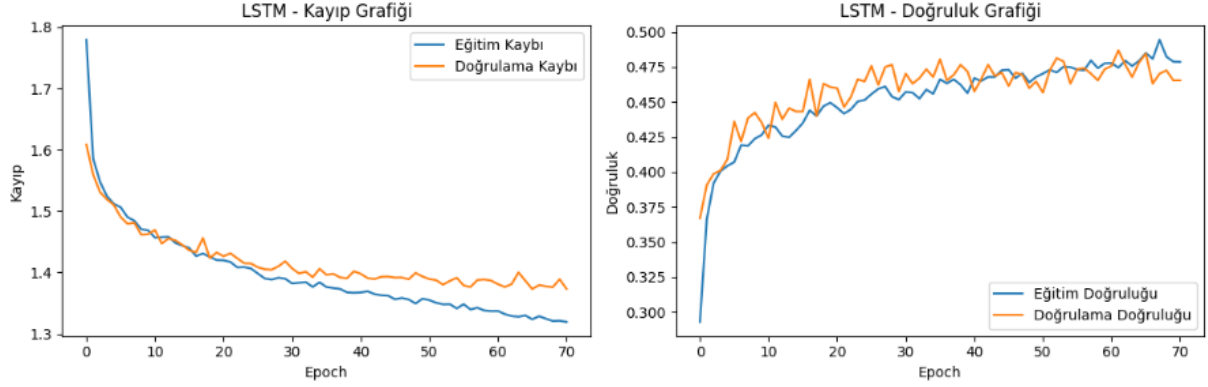
2.3.2 Model Geliştirme ve Değerlendirme

Word2Vec yöntemi için çeşitli yöntemler ve optimize edilmiş modeller denenmiştir. Optimizasyon yöntemlerinde model mimarisi değiştirilmiş ve Regularization, Batch normalizasyon gibi yöntemler denenmiştir. Aşağıda bu yöntemleri ve çıktıları yer almaktadır.

Kısmi Ön işleme + LSTM:

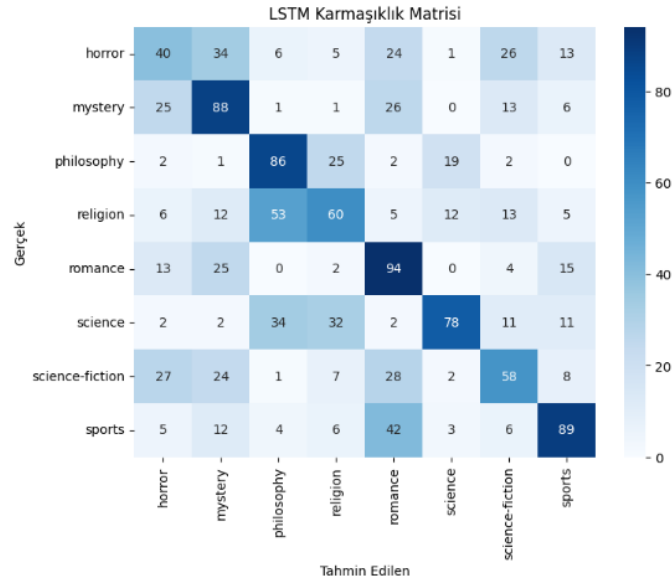
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
LSTM + Word2Vec	0.4710	0.4817	0.4710	0.4676

Grafikler:



Model performansı çok düşük. Underfit(öğrenememe) söz konusu

Karmaşıklık Matrisi:

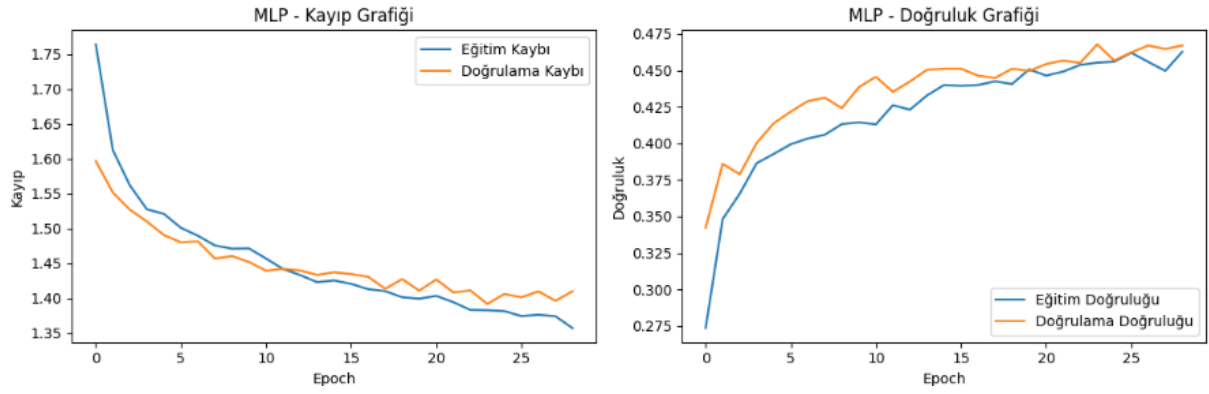


Model öğrenemediği için çoğu sınıfı yanlış tahmin ediyor.

Kısmi Ön işleme + MLP:

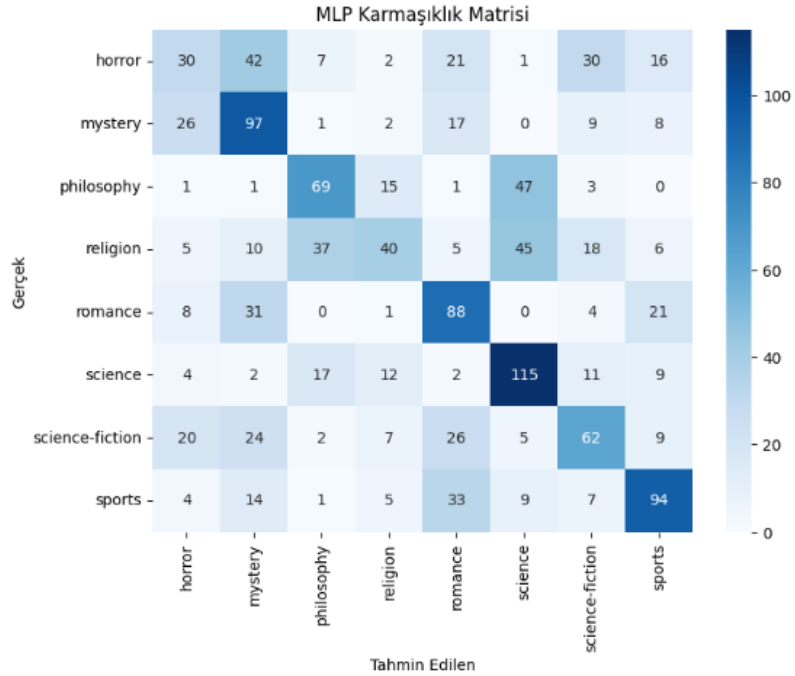
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
MLP + Word2Vec	0.4726	0.4817	0.4726	0.4593

Grafikler:



Model performansı çok düşük. Underfit(öğrenememe) söz konusu

Karmaşıklık Matrisi:



Model öğrenemediği için çoğu sınıfı yanlış tahmin ediyor.

Optimizasyon için:

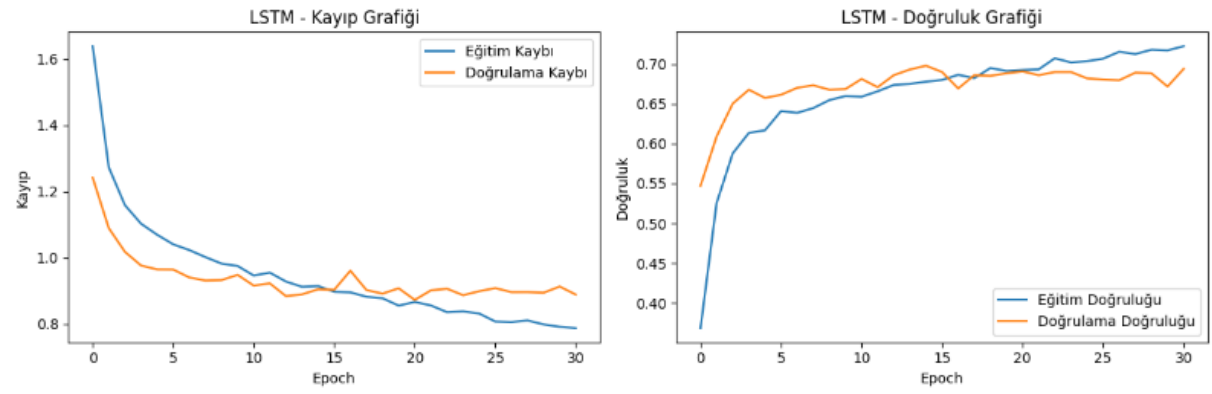
Model karmaşıklığını arttı, Regularizasyon teknikleri eklendi, Word2Vec vektörleştirmesini iyileştirildi, Eğitim parametrelerini ayarlandı

Kısmi Ön işleme + Optimize LSTM:

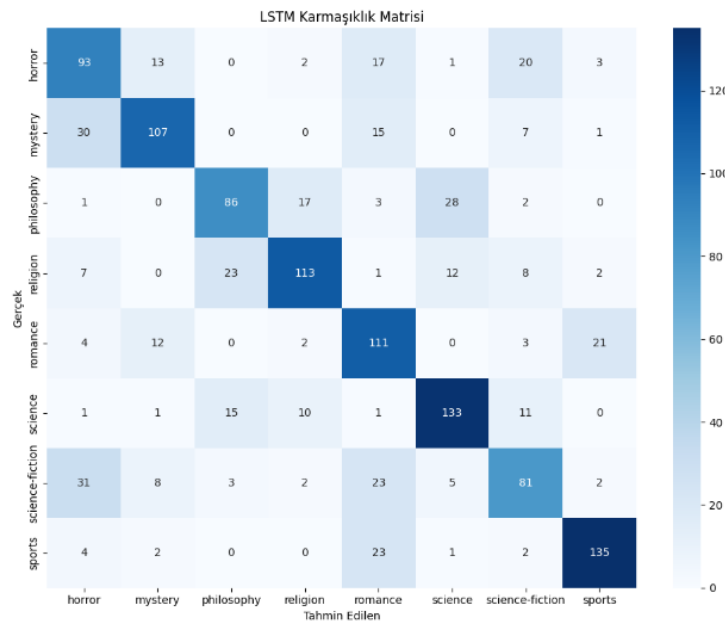
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(LSTM) + Word2Vec	0.6823	0.6893	0.6823	0.6831

Optimizasyon sonrası modelin performansı önemli ölçüde arttı.

Grafikler:



Karmaşıklık Matrisi:



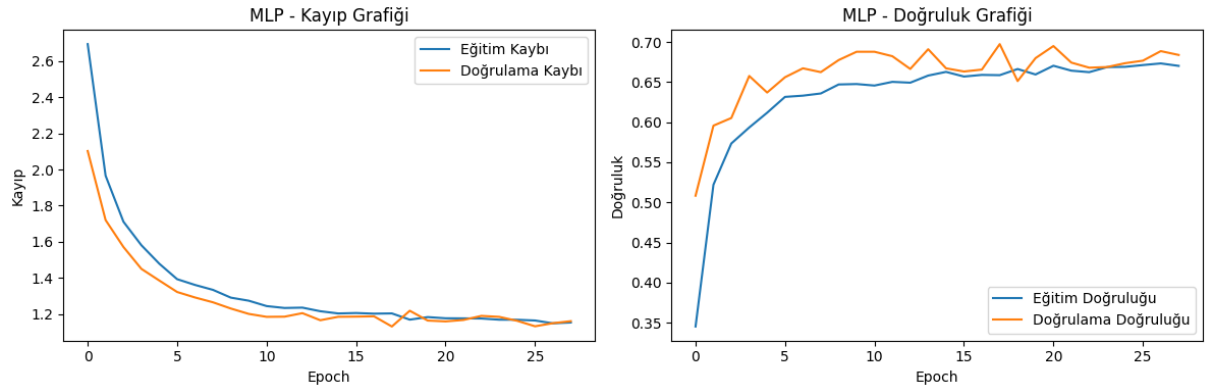
Model daha iyi bir sınıflandırma gösteriyor. Bazı sınıflarda yine karışıklık söz konusu. Horror- science fiction ya da horror-mystery, science-philosophy gibi.

Kısmi Ön işleme + Optimize MLP:

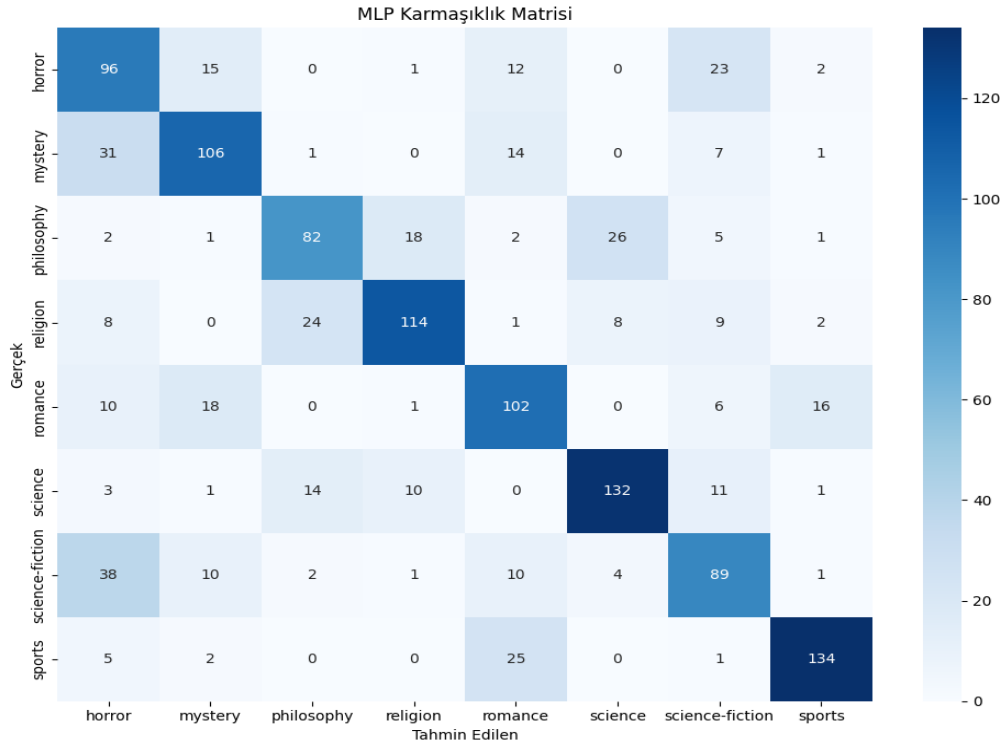
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(MLP) + Word2Vec	0.6791	0.6889	0.6791	0.6820

Optimizasyon sonrası modelin performansı önemli ölçüde arttı.

Grafikler:



Karmaşıklık Matrisi:



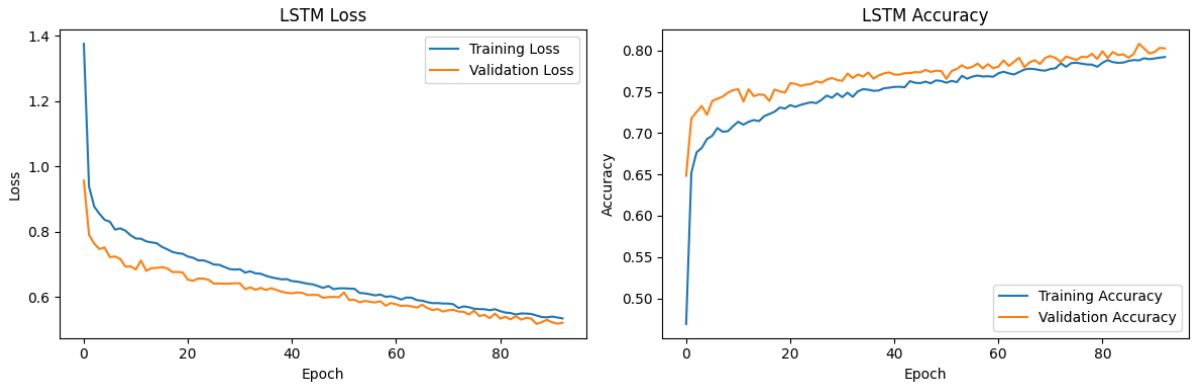
Modelin performansı arttı. Ama model belirli sınıfları karıştırıyor. Horror-Science fiction, horror-mystery gibi.

Kısmi Ön işleme + Veri Arttırma +Optimize LSTM:

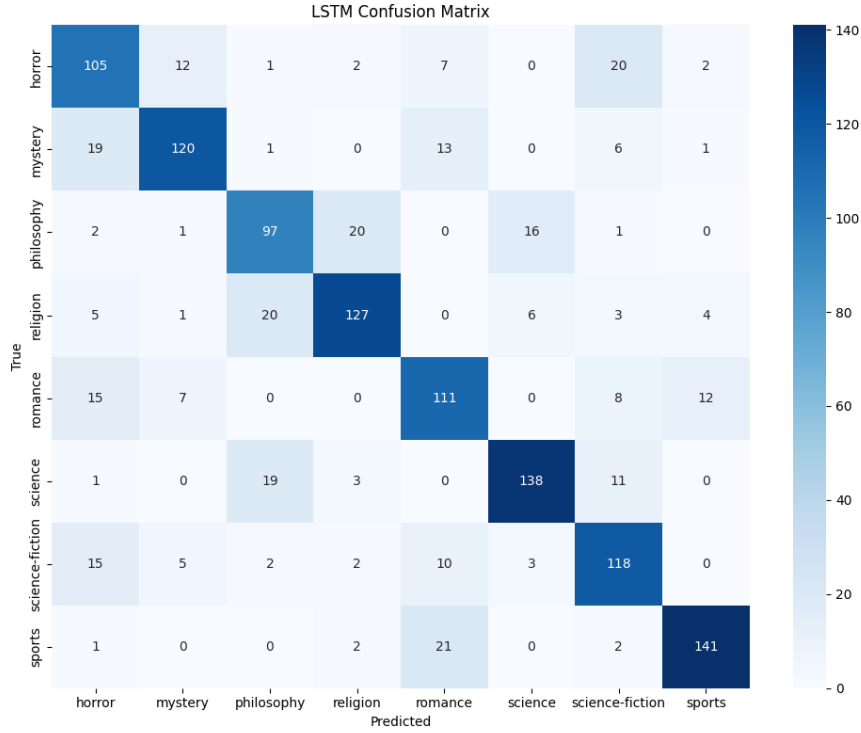
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(LSTM) + Veri Arttırma + Word2Vec	0.7601	0.7605	0.7577	0.7582

Veri arttırma modelin performansını daha da arttırmıştır.

Grafikler:



Karmaşıklık Matrisi:



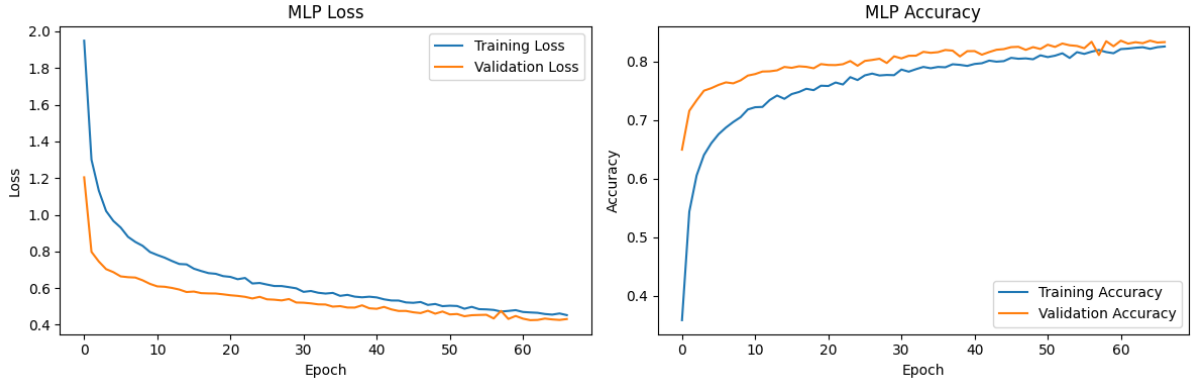
Model daha iyi genelleme yeteneğine sahip. Az da olsa belirli sınıflar karıştırılmakta. Romance-sports, horror-mystery gibi.

Kısmi Ön işleme + Veri Arttırma +Optimize MLP:

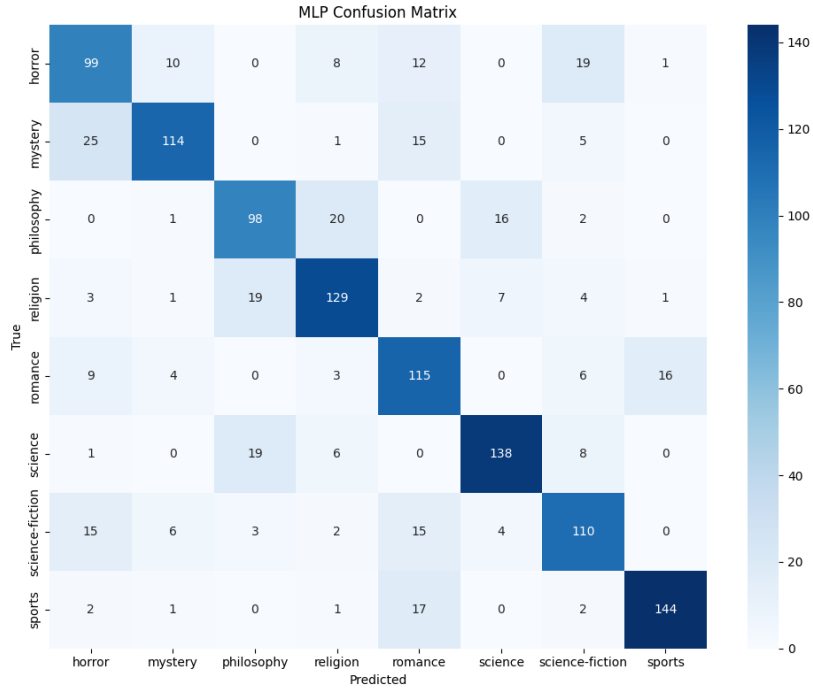
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(MLP) + Veri Arttırma + Word2Vec	0.7522	0.7528	0.7494	0.7500

Veri arttırma sonra modelin performansı önemli ölçüde arttı ve daha tutarlı bir seyreltide ilerliyor.

Grafikler:



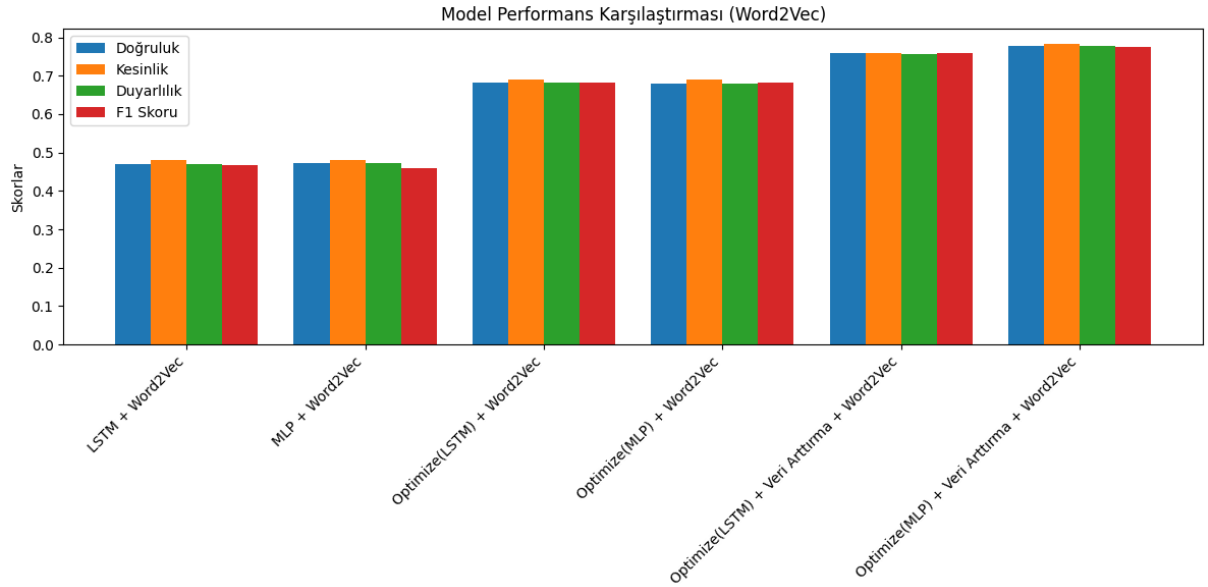
Karmaşıklık Matrisi:



Modelin genelleme yeteneği arttı. Az da olsa bazı sınıflarda karıştırılma söz konusu. Horror-mystery, religion-philosophy gibi.

Word2Vec için Nihai Tablo:

Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
LSTM + Word2Vec	0.4710	0.4817	0.4710	0.4676
MLP + Word2Vec	0.4726	0.4817	0.4726	0.4593
Optimize(LSTM) + Word2Vec	0.6823	0.6893	0.6823	0.6831
Optimize(MLP) + Word2Vec	0.6791	0.6889	0.6791	0.6820
Optimize(LSTM) + Veri Arttırma + Word2Vec	0.7601	0.7605	0.7577	0.7582
Optimize(MLP) + Veri Arttırma + Word2Vec	0.7765	0.7829	0.7765	0.7751



Genel olarak, optimize edilmiş ve veri artırma teknikleri uygulanan modellerin diğerlerine göre daha yüksek doğruluk, kesinlik, duyarlılık ve F1 skoruna sahip olduğu görülmektedir. Veri artırma tekniklerinin, optimize edilmiş modellerin performansını önemli ölçüde artırdığı gözlemlenmiştir. **Optimize(MLP) + Veri Arttırma + Word2Vec** en iyi model olarak öne çıkmaktadır.

2.4 Bert

2.4.1 Veri ön İşleme

Bert yöntemi için Bow ve TF-IDF yöntemlerinden farklı ön işleme yöntemleri uygulanmıştır. Bert bağlamsal anlamı yakalayabildiği için stop words'ler çıkarılmamıştır. Aynı şekilde kelimeler köklerine indirgenmemiştir. Noktalama işaretleri kaldırılmamıştır. Sadece aşağıdaki yöntemler uygulandığında modelin daha iyi performans verdiği gözlemlenmiştir. Veri arttırma yöntemi ile model daha da iyi performans vermiştir.

- Metinleri Küçük Harfe Dönüştürme
- Boşlukları, URL'leri, Özel Karakterleri kaldırma
- Veri arttırma

Veriler aşağıdaki veri arttırma yöntemleriyle çoğaltılmıştır.

Eşanlamlı Değiştirme (Synonym Replacement): Bu yöntem, metin içindeki kelimeleri eşanlamlılarıyla değiştirerek veri setini zenginleştirir. Modelin, farklı kelime kullanımlarına karşı daha dayanıklı hale gelmesini ve genelleme yeteneğinin artmasını sağlar.

Rastgele Silme (Random Deletion): Bu yöntem, metin içindeki kelimeleri rastgele bir olasılıkla silerek veri setini çeşitlendirir. Modelin, eksik veya gürültülü verilere karşı daha sağlam hale gelmesini ve genelleme yeteneğinin artmasını sağlar.

2.4.2 Model Geliştirme ve Değerlendirme

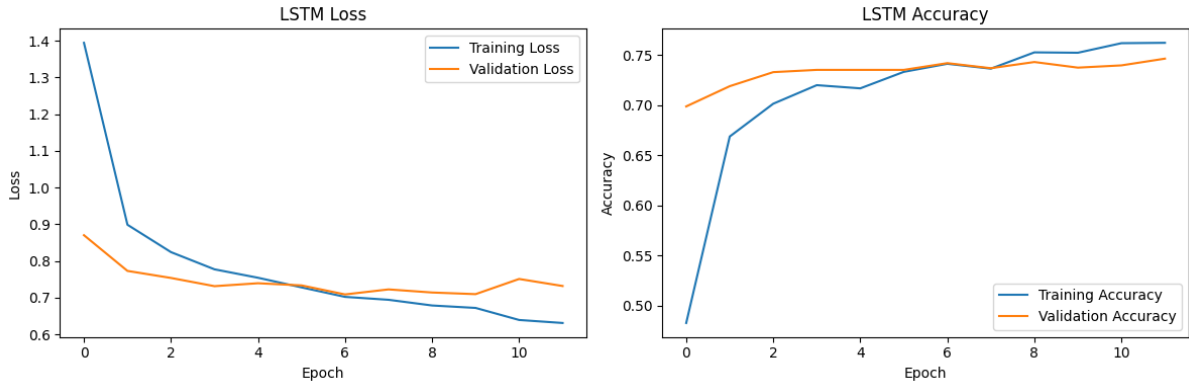
Bert yöntemi için çeşitli yöntemler ve optimize edilmiş modeller denenmiştir. Optimizasyon yöntemlerinde model mimarisi değiştirilmiş ve Regularization, Batch normalizasyon gibi yöntemler denenmiştir. Aşağıda bu yöntemleri ve çıktıları yer almaktadır.

Kısmi Ön işleme + Optimize LSTM:

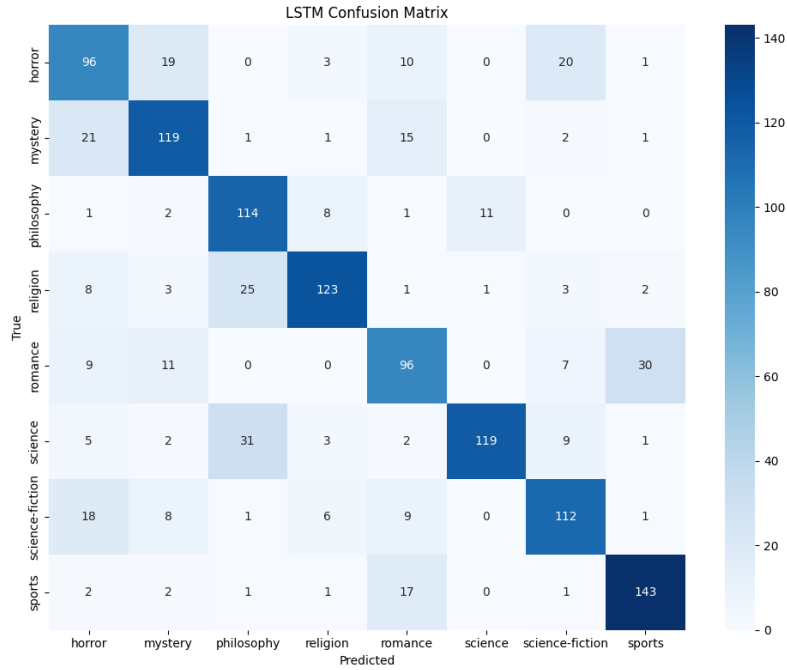
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(LSTM) + Bert(base-uncase)	0.7323	0.7368	0.7324	0.7308

Model orta seviyede bir doğruluk göstermektedir. Grafiklere bakılınca model ezberlemeye başlıyor. Erken durdurma ile az da olsa önüne geçildi.

Grafikler:



Karmaşıklık Matrisi:



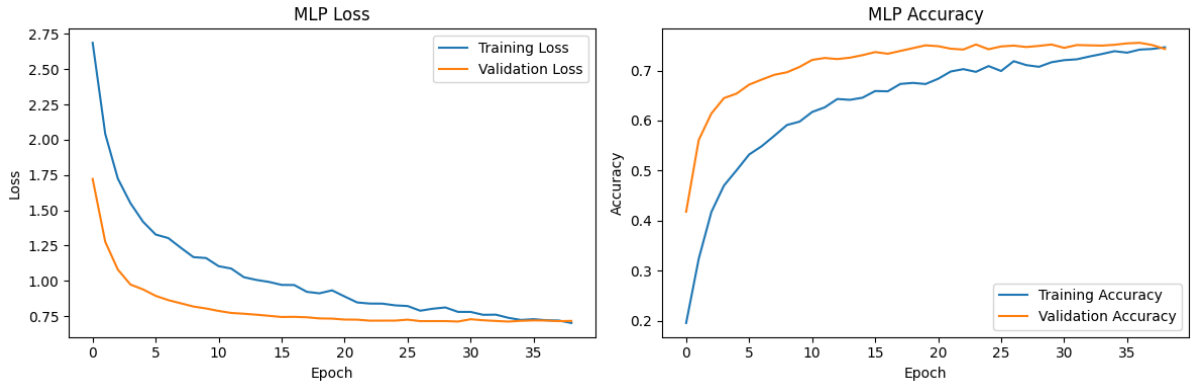
Modelin bazı türleri oldukça iyi tahmin ettiği, bazılarını ise karıştırdığı görülüyor. Bazı türler birbirine çok yakın olduğu için hata yapması beklenebilir (örneğin "Romance" ve "Sports" arasında bazı yanlış tahminler var).

Kısmi Ön işleme + Optimize MLP:

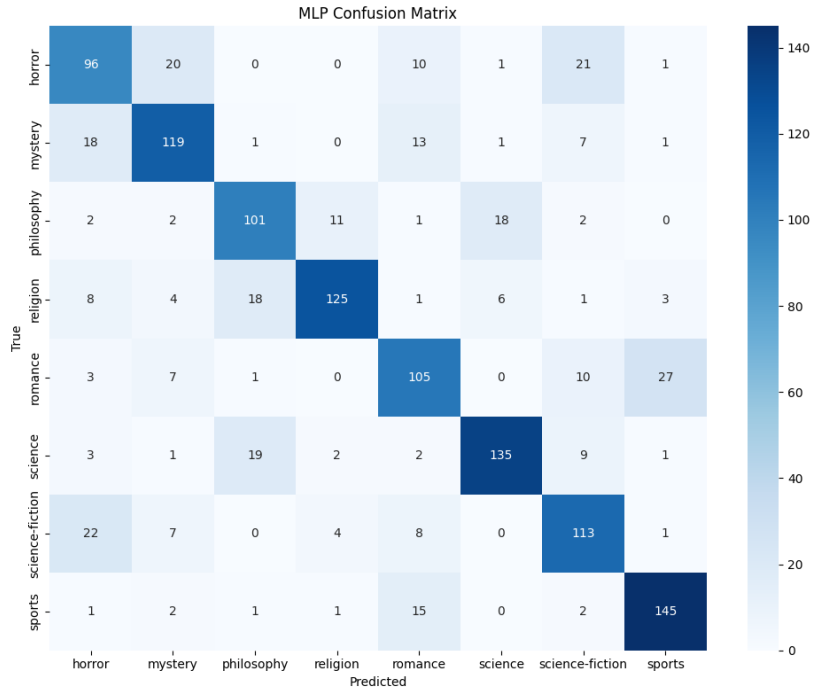
Model ve Vektörleştirme	Doğruluk (Accuracy)	Keskinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(MLP) + Bert(base-uncase)	0.7458	0.7454	0.7433	0.7435

LSTM modeline kıyasla çok az farkla daha iyi bir performans göstermektedir. Ama ciddiye değer bir fark yok

Grafikler:



Karmaşıklık Matrisi:



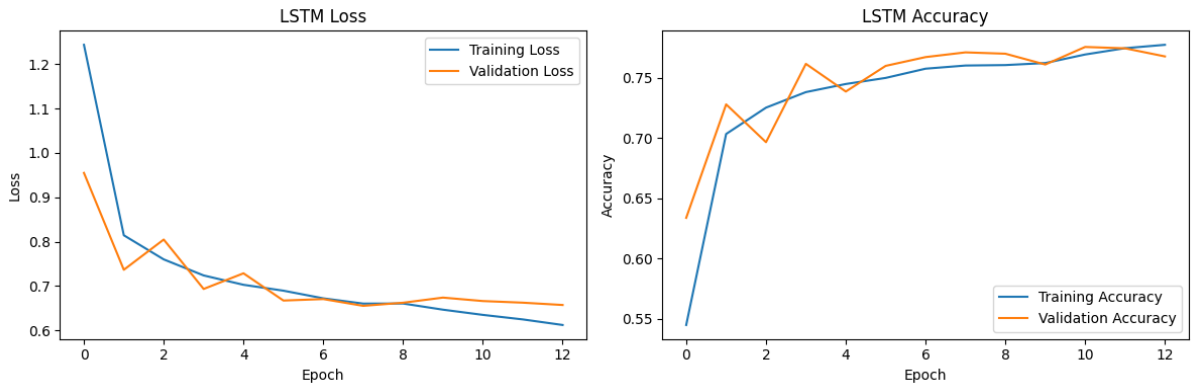
Model bazı türleri karıştırıyor. Romance-sports gibi ya da science-psilohopy gibi

Kısmi Ön işleme + Optimize LSTM:

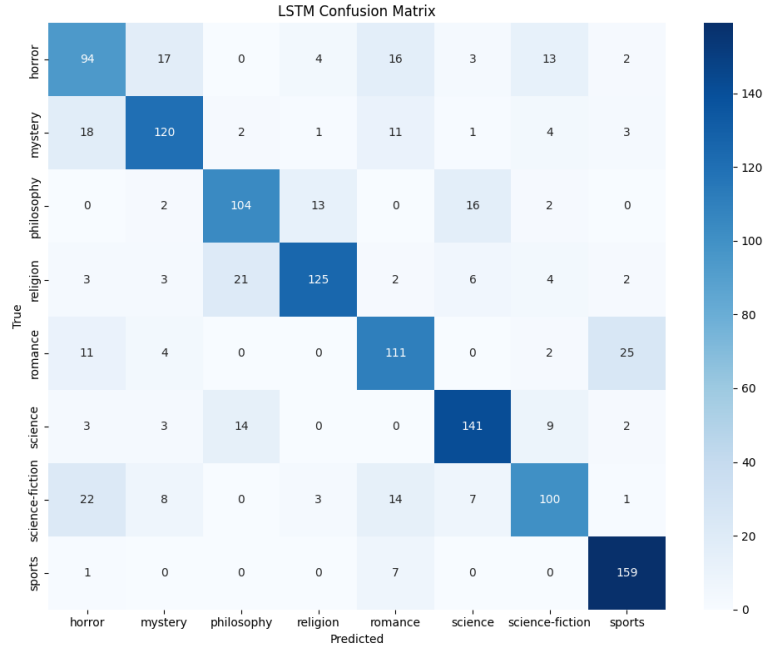
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(LSTM) + Bert(distilbert)	0.7577	0.7553	0.7544	0.7533

Distilbert modeli ile model daha iyi bir performans göstermiştir.

Grafikler:



Karmaşıklık Matrisi:



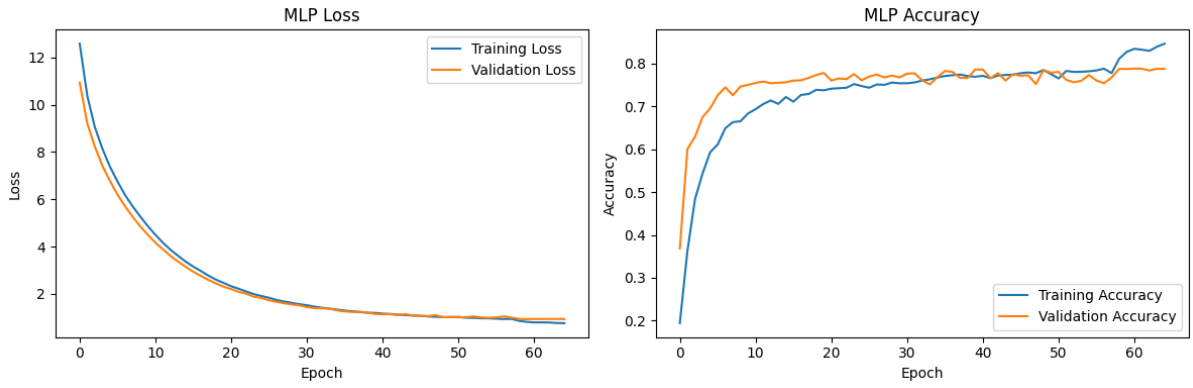
Model daha iyi bir performans gösterse bile yine bazı sınıfları karıştırıyor. Bu sınıfların en başında sports-romance gelirken hemen arkasından horror-science fiction geliyor

Kısmi Ön işleme + Optimize MLP:

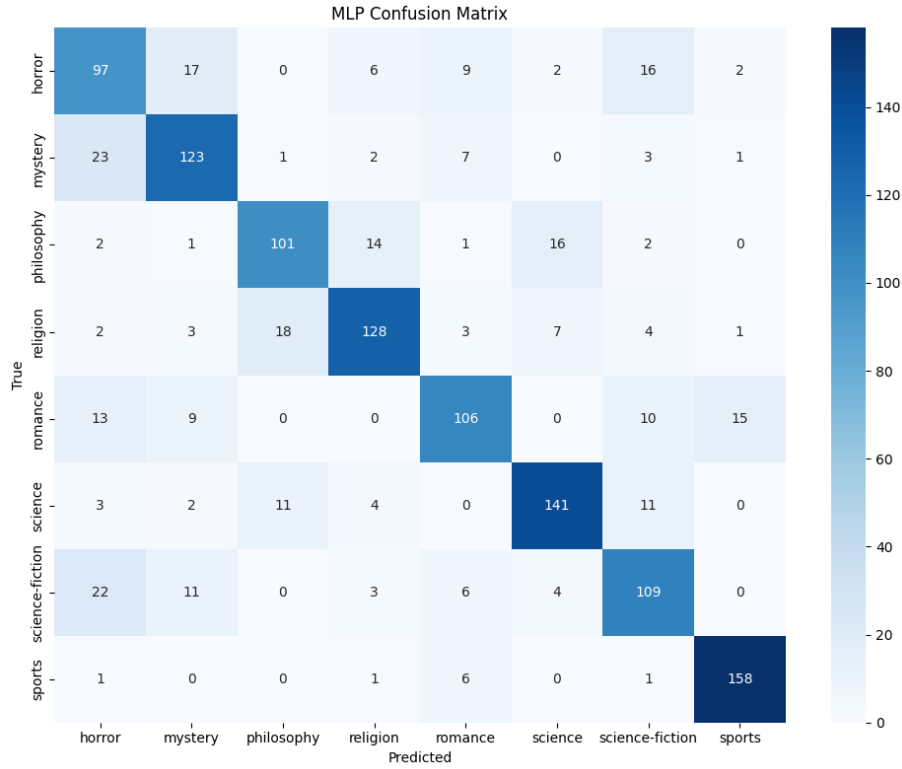
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(MLP) + Bert(distilbert)	0.7649	0.7633	0.7612	0.7616

Distilbert modeli ile model daha iyi bir performans göstermiştir.

Grafikler:



Karmaşıklık Matrisi:



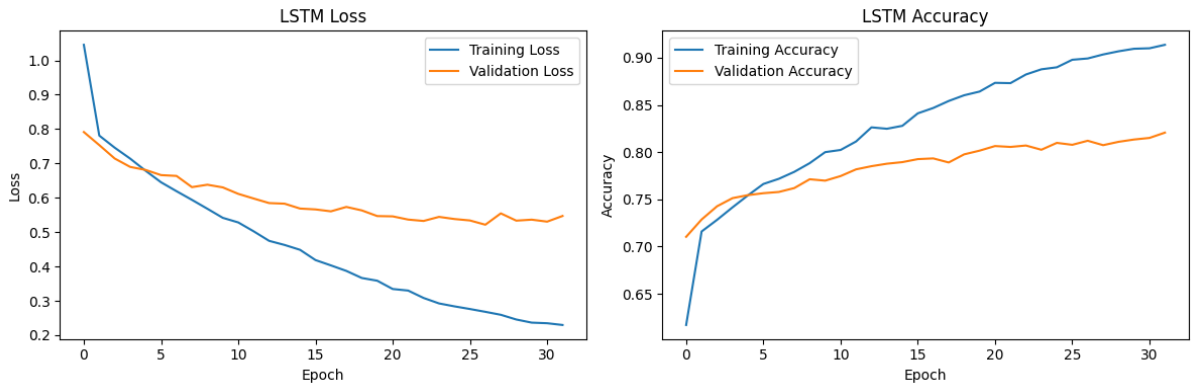
Model daha iyi bir performans gösterse bile yine bazı sınıfları karıştırıyor. Lstm modelinden farklı olarak mlp modeli ilk başta horror-mystery sınıfını en fazla karıştırmıştır.

Kısmi Ön işleme + Veri Arttırma +Optimize LSTM:

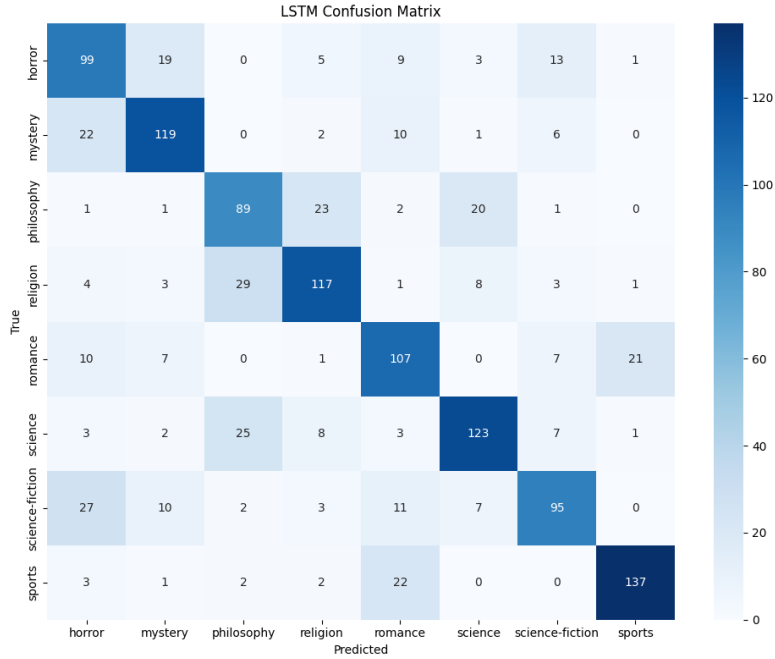
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(LSTM) + Veri Arttırma + Bert(base-uncase)	0.7037	0.7039	0.7013	0.7014

Model orta seviyede bir doğruluk göstermektedir. Veri arttırma modelin performansını düşürmüştür. Grafiklere bakılınca model ezberlemeye başlıyor. Erken durdurma ile az da olsa önüne geçildi.

Grafikler:



Karmaşıklık Matrisi:



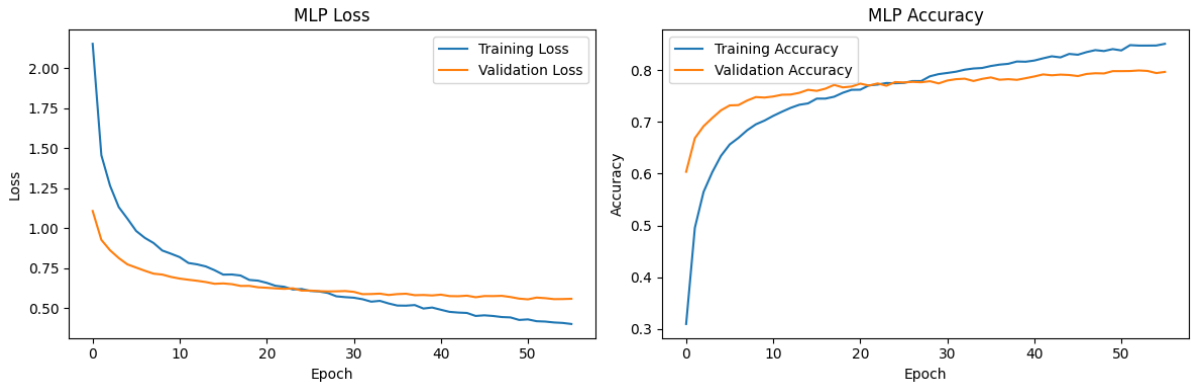
Modelin bazı türleri oldukça iyi tahmin ettiği, bazılarını ise karıştırdığı görülüyor. Bazı türler birbirine çok yakın olduğu için hata yapması beklenebilir (örneğin "Horror" ve "Mystery" arasında bazı yanlış tahminler var).

Kısmi Ön işleme + Veri Arttırma +Optimize MLP:

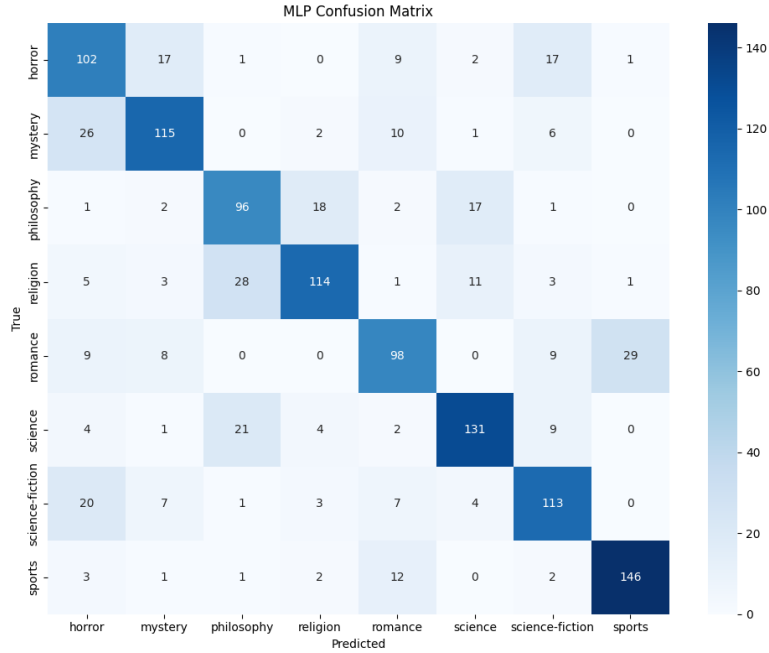
Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(MLP) + Veri Arttırma + Bert(base-uncase)	0.7268	0.7260	0.7245	0.7240

LSTM modeline kıyasla performansı daha iyi ve grafiğe bakılınca daha tutarlı doğruluk düşük olsa da overfitting çok daha az. Veri arttırma modelin performansını düşürmüştür.

Grafikler:

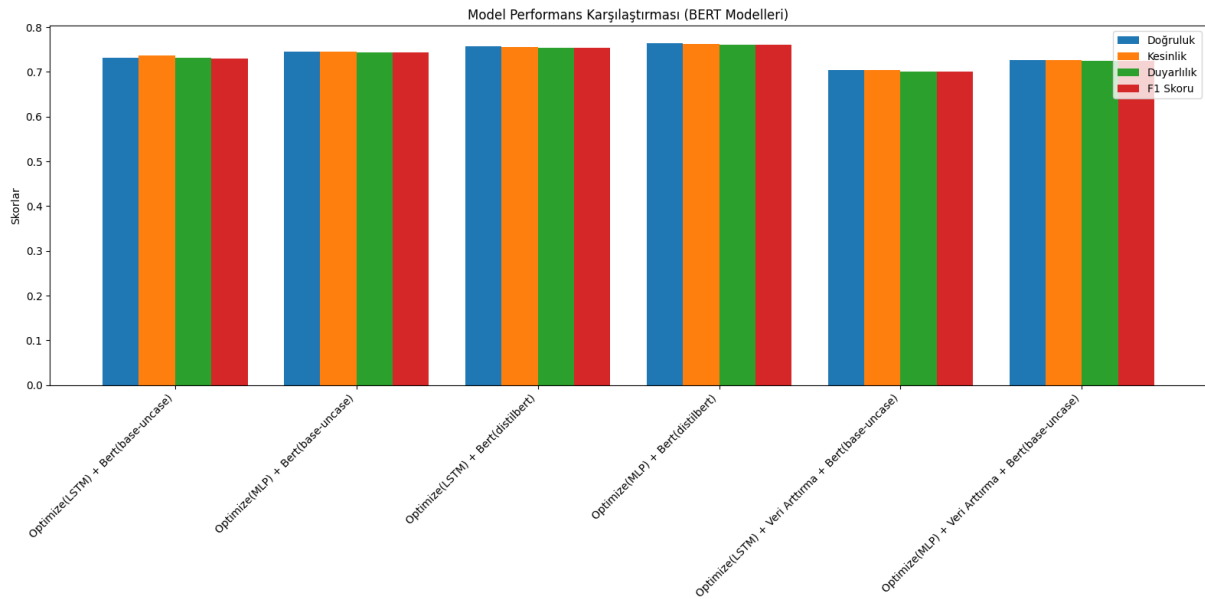


Karmaşıklık Matrisi:



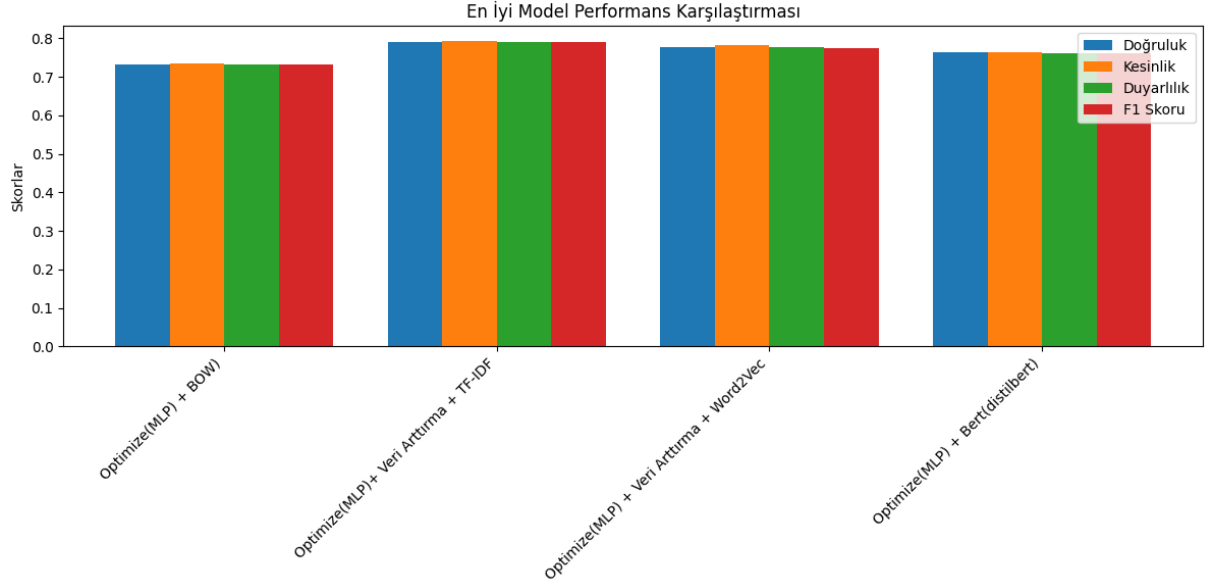
Model sports-romance, religion-philosophy gibi sınıfları karıştırmakta.

Model ve Vektörleştirme	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru
Optimize(LSTM) + Bert(base-uncase)	0.7323	0.7368	0.7324	0.7308
Optimize(MLP) + Bert(base-uncase)	0.7458	0.7454	0.7433	0.7435
Optimize(LSTM) + Bert(distilbert)	0.7577	0.7553	0.7544	0.7533
Optimize(MLP) + Bert(distilbert)	0.7649	0.7633	0.7612	0.7616
Optimize(LSTM) + Veri Arttırma + Bert(base-uncase)	0.7037	0.7039	0.7013	0.7014
Optimize(MLP) + Veri Arttırma + Bert(base-uncase)	0.7268	0.7260	0.7245	0.7240



Genel olarak, optimize edilmiş modellerin (Optimize(LSTM) ve Optimize(MLP)) ve distilbert modelinin diğerlerine göre daha yüksek doğruluk, kesinlik, duyarlılık ve F1 skoruna sahip olduğu görülmektedir. Veri artırma tekniklerinin, base-uncased modelinin performansını düşürdüğü gözlemlenmiştir.

3.Sonuç



Grafiği incelediğimizde, TF-IDF ve Word2Vec modellerinin diğerlerine göre daha iyi performans gösterdiğini görmekteyiz. En iyi değerleri **Optimize(MLP)+ Veri Arttırma + TF-IDF** yöntemi göstermektedir.

3.1. Gelecek Çalışmalar İçin Öneriler

- Daha büyük ve dengeli bir veri seti kullanımı,
- Modelin farklı optimizasyon teknikleri ile daha da iyileştirilmesi,
- Transformer tabanlı daha gelişmiş modellerin incelenmesi,
- Veri ön işleme aşamalarının daha detaylı şekilde ele alınması.
- En iyi modelleri birleştirerek voting yöntemi ile sonuç üretmesi.